

Against short-termism: priorities for a new government

Whichever political party wins power in the United Kingdom, it will need urgently to ensure that high-technology industry, investors and government develop capacities to act more strategically.

Like many countries in the West, the United Kingdom looks admiringly at the growth rates of economies in the Far East and perceives a domestic need for skilled workforces and strong value-adding innovation in services and manufacturing. Those tigerish economic statistics partly reflect a number of factors that established Western economies cannot hope to emulate, such as early stages of economic maturity in which high growth is easier to achieve, and a work ethic and national cohesion based on 'Asian values' that Westerners would find repressive.

But other factors are there too, not inextricably tied to the above, to which Western governments, industries and investors need to respond better than they have. Underlying them is a capacity to develop a technologically sophisticated long-term view — and to act on it. True, there is no lack of technological sophistication in the West, nor indeed of strategic ideas. The focus for a new government has instead to be on the obstacles to deploying them in practice.

Encouraging industrialists to invest in research and development is one factor in a long-term strategy. The international trend here is against all-embracing tax incentives. Countries that still apply such policies (of which Australia is a notable example) are tending to reduce the level of such support. Science policy research has yielded mixed evidence as to the success of such incentives. And a downward trend has a lot to be said for it. There is much anecdotal evidence from the United States, for example, that corporate technologists receive little of the money released by tax breaks, and that innovation is thereby fostered in one activity only: corporate accounting. Without imaginative ways of preventing such abuse, for a new UK administration to reintroduce such incentives — as has been suggested in Labour-associated policy circles (see page 314) — would probably prove to be an expensive substitute for meaningful action.

Frustrations

One important measure of a company's willingness to take a long-term view is research and development expenditure as a proportion of its sales. The most recent set of international statistics, published last year, indicates that, in this respect, major UK businesses are failing dismally to match the competition, pharmaceuticals apart. But less publicly noticeable are smaller high-technology companies, with turnovers in the 10s and 100s of millions of pounds, whose chief executives are likely to be technologists rather than accountants, and are all the more capable of a long-term view. Frustratingly, they too find it hard to persuade their investors that, for example, a drop in pre-tax profits is required, in order to develop in-house skills and technology whose benefits are not fully foreseeable but which, in their view, will be required to protect competitiveness in the long term. A look at their competitors in the Far East reveals all too often more sympathetic investors, and governments more willing to pump-prime new technologies.

The main response of the United Kingdom's Conservative

government is partnership with academics in several forms, especially in a continuation of the Technology Foresight programme. The commendable vigour in that gigantic exercise in industrial/academic networking, and its more questionably strong influence on research funding in universities, makes the industrial research and development statistics all the more embarrassing for the Conservatives. True, investors are more and more involved in the networks. But the problem of under-investment is too deep-seated to be solved by talking, and is likely to be cracked only by a number of small but effective changes to the investment environment — reduced taxes on capital gains made from company shares held over extended periods, for example.

For smaller companies, a growth in corporate venturing — strategic partnerships in which a large high-technology company invests in new small enterprises in the hope of synergy in the long term — is a welcome trend. With other venture capitalists, a major obstacle, particularly with successful small companies wishing to grow bigger, is one of conflicting interests: too much control and predictability sought by the investors, too much independence demanded by the industrialists.

Science too

Such investment hurdles should loom large in any new government's consideration of the health of the UK high-technology economy. But short-termism has been alarmingly prevalent also in the management of the science base. One example is an inability to implement a strategy for investment in major components of scientific infrastructure, either nationally or multilaterally. To complain, as research council chiefs do, that science funding is both inflexible and stagnant is to avoid a central issue. The science establishment too has to face up to the need to be strategic: to prioritize over a period of years in the certain knowledge that some excellent science will be sacrificed in the process. (One can accordingly expect resistance to suggestions that the advice to the government of the director general of the research councils should be published.)

A second significant problem of government short-termism lies within many of its departments, most notably in another major source of government embarrassment, the Ministry of Agriculture, Fisheries and Food (MAFF). Talk to scientists funded by it, and you hear of funds switched on and off like a tap (see also page 316). That short-termism is a symptom of wider problems in that department. The Labour party has promised to separate consumer protection responsibilities from agricultural industry interests — a welcome but deplorably belated development. That change will not go deep enough, however. Under a new administration, the strength of the Office of Science and Technology will need to be increased in the teeth of political opposition in order, not least, that MAFF's internal networks of science advice and funding, in addition to the quality of its laboratories' research, are subject to adequate external scrutiny. □



US academy warns of disaster if required to open up panels

[WASHINGTON] An appeal court ruling expected as early as next week will help to decide whether the US National Academy of Sciences (NAS) and its associated organizations, which provide advice to the US government on scientific and technical matters, is subject to public access rules known as the 'sunshine laws'.

The rules are known officially as the Federal Advisory Committee Act (FACA), passed in 1988, and are intended to open up the government advisory system to closer public scrutiny. But officials in the various organizations are concerned that applying FACA to their work would undermine their independence, allow the government to exert direct control over committees, and wreck a carefully constructed peer-review process.

"This is a very serious issue for us," says William Colglazier, executive director of the academy. Under FACA, he says, government officials would have to approve the membership of committees — which must include non-experts. And the review process would be disallowed, because its oversight of the public committees would be illegal. "The key thing for us is our independence," he says.

The academy complex, which includes the National Academy of Engineering and the Institute of Medicine and their operating arm, the National Research Council (NRC), prides itself on its independence from the government agencies that request and pay for its reports.

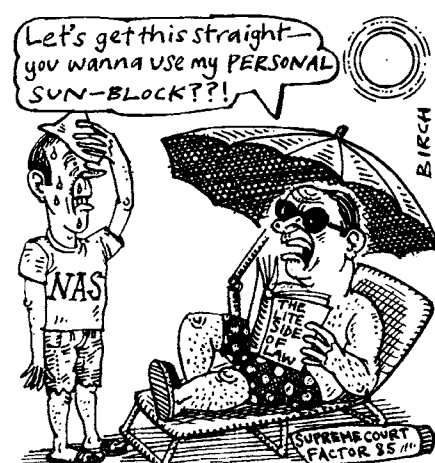
But this way of operating came under legal threat in January, when three judges at the US Court of Appeals for the District of Columbia accepted the argument of the Animal Legal Defense Fund (ALDF) that the academy is a 'quasi-public organization' and therefore subject to FACA.

The court based its judgement on a Supreme Court case in 1989 which determined the status of the American Bar Association under FACA. In its ruling, the Supreme Court cited the NAS as an example of a non-government group that should be subject to FACA. NAS was not involved in the case, and chose not to appeal against the ruling.

Last week, after a separate court case, the academy was obliged to release a report on the proposed National Ignition Facility under unusual, court-imposed conditions that preclude its sponsor, the Department of Energy, from making use of its findings (see *Nature* 385, 755 & 386, 210; 1997).

In a statement accompanying the report, Bruce Alberts, president of the academy, said: "We strongly maintain that committees of the academy and its sister organizations do not fall subject to FACA. Were we to fall subject to FACA, we would be compromised severely in our ability to provide independent advice — as we have done, in the national interest, for more than 100 years."

While the academy appeals against the January ruling, officials at the 1,500-staff complex are closely studying its possible consequences. "We're looking at how we



could operate within FACA," says Colglazier. That could mean asking individuals to produce reports as 'principal investigators', consulting whoever they see fit, he says. Colglazier also confirms that bemused NRC staff have been asked to stop using the word 'committee' in their communications with sponsoring agencies.

Tom Grumbly, under-secretary for energy, alluded to the severity of the crisis at the complex in testimony last week before an appropriations subcommittee of the House of Representatives. "The issue of the role of the NRC and whether it is subject to FACA is a big deal right across the government," he said. "I know the academy has a very strong view that [compliance] will make the role of the NRC very difficult."

Environmental groups which have sued the academy complex believe that it is insufficiently independent of government and that its mode of operation is elitist and exclusive. But a spokeswoman for the academy says that the NRC was set up partly to ensure that the process was inclusive, and that four-fifths of its panel members are now drawn from outside the elite academies themselves.

The ALDF last week filed documents with the full US Court of Appeals for the District of Columbia in response to NAS's request for an appeal hearing. Its move has raised expectations at the academy that the court could announce as early as next week if it is prepared to hear the academy's appeal on the January case — and if it will rule that the January decision is void while an appeal is heard.

If the court rejects the appeal, the academy faces a period of upheaval while it goes either to the Supreme Court for a final ruling or to Congress for a change in the law to exempt it from FACA.

Colin MacIwain

Russian science minister faces challenge

[MOSCOW] Science has been given back a seat at the cabinet table of Russian President Boris Yeltsin in a series of ministerial reforms announced last week. But the incoming science minister faces an unusual challenge — he will have to work with a new deputy prime minister charged with reviewing Russia's science.

The new science ministry replaces the old State Committee for Science and Technology, and Vladimir Fortov, the committee's chairman, has been given the job of science minister.

But controversy surrounds the responsibilities that accompany Fortov's new job. The task of reorganizing Russian science has been given to Vladimir Bulgak, the former communications minister, who has been made a new deputy prime minister.

There is also uncertainty about Fortov's new ministerial status. Although he has



Fortov: minister, but demoted?

been promoted from chairman of the state committee to full cabinet minister, Yeltsin's decision to relieve Fortov of his post as one of eight deputy prime ministers is being interpreted by some as a demotion.

Some observers believe that Bulgak has been brought in to curb Fortov's enthusiasm for the Russian Academy of Sciences.

Throughout his tenure as chairman of the state science committee, Fortov had tried to raise the academy's influence over policy-making, to the dismay of some seeking broad structural reforms of Russian science.

Bulgak's former communications ministry has been reconstituted and named as the State Committee for Communications and Information.

Ulster violence leads to university cuts

[MUNICH] The resumption of violence in Northern Ireland is having a major impact on its two universities, as money to develop research facilities is being cut to help pay for extra security measures in the province.

For the next academic year, which starts in August, the Northern Ireland Higher Education Funding Council (NIHEFC) is planning to reduce by 16 per cent its core support for research at the Queen's University of Belfast and the University of Ulster. The council's provisional plan is that research funding to the two universities, which is now £25 million a year, will be cut by 25 per cent in the following two years.

The universities claim that the cuts will mean up to 200 job losses, and that they will have a particular impact on young researchers currently employed on short-term contracts.

The cuts are being made by eliminating part of the government grant to universities which is known as Development Research — or DevR — money. This has been provided by the funding council since 1992, when to have allocated funds solely on the results of the first Research Assessment Exercise would have resulted in the two universities facing large reductions in their income.

DevR money has been selectively allocat-

ed to a number of universities, both in Northern Ireland and on mainland Britain, to help them build up their research competence. In Northern Ireland, this extra money has made it possible to maintain the funding council's support for research at its 1992 level.

Scientists feel particularly aggrieved because the results of the second Research Assessment Exercise, carried out jointly by the four UK university funding councils, and published in January (see *Nature* 385, 3; 1997), showed that the extra money had been well spent. Both universities showed considerable improvements in the quality of their research in particular disciplines.

Biomedical research at the University of Ulster received 5* — the top score — in the most recent assessment exercise, compared to 4 four years ago. At Queen's, the number of departments placed in the top three categories (out of seven) increased from eight to 20.

The universities say the cuts will erode the research infrastructure, making it difficult to attract money from other sources. Bob Cormack, vice-chancellor at Queen's, says that "for every pound of DevR money spent, we are able to raise £1.25 from other sources". The universities are demanding a reversal of the claw-back, and local industry, with which the universities enjoy a close working relationship, is backing the demand.

Some are pessimistic about their prospects. Eric Beatty, director of the Northern Ireland Technology Centre at Queen's, argues that it will be difficult to persuade the government to change its mind. "Every sector is feeling the consequence of increased security spending since the restarting of the violence."

Beatty says the threat of even bigger cutbacks in 1999 and 2000 has a greater chance of being averted, because high industrial support for the universities — which provide the only significant research base in the province — is likely to influence future budget discussions.

But others are more optimistic about the shorter term. An independent review of the spending of DevR money, commissioned by the funding council from the consulting group Segal Quince Wicksteed, based in Cambridge, is due to be published in the next couple of months.

A spokesman for the University of Ulster believes that the report may have a significant influence, when combined with "demands from all over industry for reversal of the cuts". Peter Holmes, under-secretary for universities at the funding council, hints that this may, to some extent, be the case, suggesting that the government may decide to "temper" the cuts if the review turns out to be highly positive about the way the money has been spent.

Alison Abbott

Consensus grows on risk assessment

[LONDON] A series of events over the past year — including the crisis over 'mad cow disease' and the proposed dumping of the Brent Spar oil platform — appears to have triggered a growing convergence between the views of UK scientists and social scientists on the handling of risk.

That was the message to emerge from a meeting at the Royal Society in London last week. Five years ago, a similar meeting ended in bitterness and hostility, and generated a report in two contradictory parts. The scientists maintained that risks can be 'objectively' measured, while the social scientists insisted that risks are 'culturally constructed'.

But at last week's meeting there appeared to be consensus on the principle that public perceptions must be included in the assessment of risks. "There will never be a universal Richter scale of risk," admitted Derek Burke, former vice-chancellor of the University of East Anglia, and chairman of the UK Advisory Committee on Novel Foods and Processes.

John Ashworth, a former chief scientific adviser to the government and the organizer of the meeting, said afterwards that a "dialogue of the deaf" appeared to have been replaced by a willingness to listen. "There is a different atmosphere from the early 1990s, and some things that were said then are not now apparent," said Ashworth. "The debate has advanced."

John Adams, a geographer at University College London, gave one reason why public perceptions are important in developing regulatory policies. "The purpose of measuring risk is to inform behaviour, which then alters that which has been measured," he explained.

As an example, Adams pointed out that, although the number of road deaths has fallen dramatically as a proportion of the population since the beginning of the

century, parents perceive the risk to children to be much higher. As a result, they limit their children's exposure to traffic, and the number of fatalities therefore falls.

Several of those attending the meeting suggested that issues such as bovine spongiform encephalopathy and Brent Spar appeared responsible for the turnaround in opinion. "When [our advisory committee] started, we did not grasp the very different way the consumer sees risk," says Burke, a difference that resulted in scientists being seen as "arrogant, distant and uncaring".

Angela Wilkinson of Shell UK suggested that the company had learnt from its mishandling of the Brent Spar controversy. "It's not a matter of cold logic," she said. "Perceptions are realities that reflect values."

Yet, despite agreement on the importance of public perception, differences remained over how this should be taken into account. John Krebs, chief executive of the Natural Environment Research Council, said that "values and perceptions should be part of science", and suggested that the social and technical sciences should be integrated to arrive at a numerically quantifiable measurement of risk.

But others saw an unwillingness among scientists to appreciate the reasoning, however 'unscientific' it might seem, behind public sensibilities. In the case of Brent Spar, for example, the public was concerned about the precedent that would be set by allowing the oil platform to be dumped at sea.

According to Robin Grove-White, director of the Centre for the Study of Environmental Change at the University of Lancaster, risk assessments are too narrow and reductionist, focusing on individual cases to the neglect of wider issues. Together with other social scientists present, he advocated greater public participation in decision-making as a way of raising the level of public trust in scientists.

Ayala Ochert

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Up and away: the Parkes multibeam device is hoisted towards the telescope's focus cabin.

Multibeam survey sweeps galaxies

[PARKES, AUSTRALIA] The 64-metre telescope at Parkes, Australia, began last week the first 'blind, all-sky' survey of radio galaxies in the southern skies using a unique multi-beam device that gives a wide-angle view, greatly increases sensitivity, and will shorten the survey from several decades to a few years.

The multibeam device was commissioned last week by Australia's science minister, Peter McGauran. The cost of A\$3 million (US\$2.3 million) was augmented with about £200,000 (US\$320,000) from Britain for equipment from the Jodrell Bank telescope. The device provides simultaneous observations of 13 separate radio sources in the 21-cm wavelength of neutral hydrogen at two polarizations.

Astronomers claim that the device is "a thousand times faster than anything before and anything envisaged for the next five years". The survey will detect galaxies out to a redshift of 10,000 km per second, and its sensitivity for discovering pulsars has been increased fivefold.

Three previously unknown galaxies were found by the multibeam in its first few days, one appearing from behind the equator of the Milky Way. The project leader, Lister Staveley-Smith, says this proves the instrument's capabilities to receive signals through regions where absorbing matter is densest.

Parkes, first brought into service in 1961, has now received two major upgrades to extend its flexibility and working life. The first, financed by the US National Aeronautics and Space Administration, involved installing a larger cabin at the focus of the altitude-azimuth mechanism to allow for the fast change of receivers.

Peter Pockley

'Less bureaucracy' pledge for Italian space agency

[ROME] Italy's space agency, ASI, started operating this month under new rules which, say agency officials, will allow it to run much more smoothly. The reform follows several years of chaos and crisis during which funding was held back from space scientists while huge debts were run up to meet national and international commitments.

The new rules were approved last month by the research minister, Luigi Berlinguer. They create for the first time a science directorate within the agency. The head of the science directorate will report directly to the agency's new president, Sergio De Julio, although the heads of the technical, applications and administration directorates will report to the director general, Giovanni Scerch.

"This move was necessary because research requires a much less bureaucratic approach," says De Julio, who took office last autumn. He is a professor of engineering at Calabria University and was formerly a member of parliament.

For the past six years, Italy's space scientists have experienced lengthy delays, often lasting for years, in obtaining promised funds for research projects. Not only has this been an obstacle to the effective running of national research programmes, but it has also been embarrassing in international circles. In the European Space Agency (ESA), for example, Italy has developed the reputation of being an unreliable partner for having delayed or reneged on agreements essential to international programmes (see *Nature* 373, 467; 1995).

Since its creation in 1988, ASI has been plagued with management problems — at times giving rise to charges of corruption — and financial difficulties. But De Julio points out that ASI has made considerable progress in controlling its debts through bank loans and internal savings. One decision was not to start any new programmes in 1997. De Julio describes this as "a very painful process which I do not wish to repeat", but says it was necessary to help repair the agency's battered credibility within the Italian parliament, which has to approve its annual budgets.

The new rules create a more powerful executive within ASI, simplifying the strategic and financial planning of research. Advisory committees have previously wielded strong influence over detailed planning, while external financial controllers decided what constituted basic research — which must by law receive 15 per cent of ASI's total budget. That left scientists frustrated at decisions made.

De Julio says that the agency will focus its spending more strategically than in the past. Immediate priorities will be in areas of prior investment, such as the international space station. "Some would argue that we shouldn't have put so much into the space station," he admits. "But we now have to consider carefully how to use the percentage of free space station usage that our investment has bought." He also plans to introduce more rigorous systems for evaluating ASI-funded research and for promoting the use of research results.

De Julio's efforts have been generally welcomed in the space science community. Giancarlo Setti, for example, professor of astronomy at the University of Bologna, says that ASI should now be able to function more efficiently. In the past, he says, a lack of funding continuity has been "most damaging" for space science programmes. Setti chairs a new scientific advisory committee to the agency, which met for the first time this month.

Pietro Ubertini, head of the space group at the Institute of Space Astrophysics in Frascati, says that he, too, is "optimistic that the new structure will work, and give a strong direction to the agency". As the principal investigator of one of the instruments on ESA's next gamma-ray astronomy mission,



De Julio: aims for strategic focus.

Integral, Ubertini has had to keep the mission going, despite ASI's continual failure to supply promised money.

"We have managed with the support of our home institutes which have usually been prepared to use their budgets to tide us over temporarily," he says. "We have also been able to cope because of our inherent enthusiasm [for the project]."

With the reforms under way, De Julio is now sufficiently confident to have drawn up a five-year plan for the space agency, to start next year. The plan assumes that parliament will agree to increase the annual budget by about 10 per cent in real terms before 2000.

ESA last week confirmed the appointment of an Italian as its director general (see *Nature* 386, 104, 1997). Antonio Rodota, an electronics engineer and a graduate of Rome University who is currently director of the space division of the Italian company Finmeccanica, will succeed Jean-Marie Luton, ESA's Council announced. In addition, the contract of Roger Bonnet, director of ESA's science programme, was extended by two years.

Alison Abbott

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Anyone out there? The sequence of bottles in the exhibit at MIT matches that of the 1974 code.

Bottled radar pays tribute to Sagan

[BOSTON] Visitors entering Building 68, the new biology facility at Massachusetts Institute of Technology (MIT), will have been struck this month by an imposing row of wooden structures spanning the length of the corridor along the building's facade.

The structures consist of 18 racks housing 1,679 glass bottles, some filled with water, others empty. The sequence of the bottles replicates the sequence of radar pulses and pauses of a coded message that was transmitted into space in 1974 by astronomers Frank Drake and Carl Sagan from the Arecibo Observatory in Puerto Rico.

The message conceived by Drake and Sagan was sent in the hope that extraterrestrial listeners might decipher it and learn about the building blocks of life, the structure of DNA, the number of nucleotides in the human genome and other facts about life on the third planet from the Sun.

'A Message in Many Bottles' is the work of Joe Davis, an artist and research affiliate with MIT's Laboratory of Molecular Structure. Davis has dedicated the installation to the memory of Carl Sagan, who died last year. The MIT Council of the Arts, which deemed the work "insufficiently artistic" to receive funding on its debut in 1988, provided \$250 this time around for moving expenses.

The exhibit, which is on show until early April, pays homage to Sagan's legacy, part of which consists of the radio signal that is propagating towards the globular star cluster M13 in the Hercules constellation, 25,000 light-years from Earth.

Philip Sharp, chairman of MIT's biology department, describes the exhibit as a "fitting tribute" to Sagan's work. "It brings the abstraction of a radar message into an accessible, physical form," says Sharp. He says he sees "numerous benefits" in having an artist who approaches issues from an unorthodox perspective working alongside more formally trained scientists.

Steve Nadis

NIH is likely to challenge genetic 'probe' patents

[WASHINGTON] The National Institutes of Health (NIH) are likely to contest a decision by US patent officials to allow the patenting of expressed sequence tags (ESTs) — short sequences of DNA that uniquely identify full-length expressed genes.

Last Friday (21 March), the NIH sent a letter to the Patents and Trademarks Office (PTO), arguing that ESTs do not meet the office's own definition of practical utility. The letter also urges the PTO to make as narrow as possible the scope of EST patents that NIH believes are legitimate because the ESTs in question correspond to known genes.

NIH's director, Harold Varmus, had said earlier that he was "concerned" about a recent decision by the PTO to consider applications for such patents on a case-by-case basis (see *Nature* 385, 755; 1997).

PTO officials argue that ESTs alone can be legitimately patented because of their utility as molecular probes. The decision comes despite a continuing debate in the scientific community about the usefulness of ESTs — as well as concern that rights to one could encompass rights to the full-length gene, and discourage further research on that gene.

Reflecting broad concern among scientists about the implications of the decision, Varmus said that "this is an issue that we are likely to publicly contest". He added: "Exactly

how that kind of patent would be issued, what rights would be given to the patent holder, are not yet known." And he warned that the patenting of such research tools could produce "disincentives to develop products and possibly strong disincentives to do the research that's required". Varmus was speaking at a seminar on biomedical research at Brookings Institution in Washington DC.

Clarisa Long, a molecular biologist and attorney who is an intellectual property expert at the American Enterprise Institute, defended the PTO's decision as reversing the previously existing "legal fiction" that ESTs lacked utility unless the function of their corresponding gene was disclosed. She also challenged the claim that narrow patents on such sequences would impede further research.

But this view is contested by Rebecca Eisenberg, a professor of law at the University of Michigan, and an adviser to NIH on intellectual property issues. "A proliferation of patent rights [under which a researcher] would need licences creates a problem for downstream product development," she says, adding that patents on multiple ESTs might have that effect for products related to therapeutic proteins, or products encoded by the sequences.

Meredith Wadman

Law urged to check genetic discrimination

[WASHINGTON] Guidelines for legislation to prevent discrimination by employers against individuals on the basis of their genetic characteristics was made public last week by a consortium of US government advisers on the social and ethical implications of genetic research and breast cancer activists.

The proposals coincide with growing public concern in the United States about the implications of genetic testing for jobs and health insurance. They also follow the introduction of several bills in Congress that seek to outlaw genetic discrimination.

The recommendations have been drawn up by the National Institutes of Health/Department of Energy Working Group on the Ethical, Legal and Social Implications of Human Genome Research (the ELSI working group) in collaboration with the National Action Plan on Breast Cancer, an organization with representatives from both the government and the private sector. They follow similar guidelines on genetic discrimination in health insurance published by the same

group in October 1995.

In a paper published in *Science*, the group recommends that employers should be barred from making promotion, pay and other benefits conditional on genetic information, unless they can prove that the information is job-related and "consistent with business necessity". In such cases, they would need to get an employee's written and informed consent for collection or release of the information.

The group also proposes that employers should be banned from requiring genetic information or testing before offering a job, and banned from gaining access to genetic information in employees' medical records. Employees alleging discrimination would be able to file private lawsuits against employers.

At present, there are few legal constraints preventing US employers from requiring employees to take genetic tests. Restrictions on employers' use of 'genetic information' — a broad term that includes physical characteristics and family history — are even more scarce.

M. W.

Tokamak closure leaves a hole at centre

[PRINCETON, NEW JERSEY] Next week sees the close of a major chapter in fusion research, when for the very last time plasma races around the United States' largest experimental tokamak, at Princeton Plasma Physics Laboratory (PPPL).

Deuterium and tritium ions at up to 500 million degrees Celsius — the hottest matter in the Solar System — will then decline in temperature to the ambience of a spring day in New Jersey, leaving the surrounding laboratory facing an uncertain future.

But the 500 staff of PPPL will have little time for philosophical reflection in the coming weeks. In June, John Schmidt, the laboratory's interim director, will implement a necessarily brutal reorganization that could see as many as 200 of them lose their jobs.

Those remaining will have rapidly to establish a new direction for the laboratory to enable it to retain a leading role in magnetic fusion research without the Tokamak Fusion Test Reactor (TFTR). This machine, which cost \$1 billion to build, completes 15 years of scientifically fruitful operation next Thursday (3 April).

The embattled US magnetic fusion community, whose budget has lost almost half its spending power since 1995 (see *Nature* 375, 622; 1995), is looking to Princeton to provide such leadership after the TFTR shuts. "We expect Princeton to continue to be a centre of excellence, but on a smaller scale," says Anne Davies, head of the fusion programme at the US Department of Energy.

"Our leadership role will change," says Schmidt. "It will derive from the people we have here and their interaction with other institutions." Success, he declares, "will depend on us — not on the Department of Energy". Rush Holt, assistant director, agrees: "Leadership will be ours if we show that we can lead," he says.

To help with the transition, Princeton is being allowed to build the National Spherical Torus Experiment (NSTX), a \$25-million machine in which the doughnut shape of a conventional tokamak is sharply compressed into a ball with a narrow channel running down its centre. The machine, conceived at Oak Ridge National Laboratory in Tennessee, is due to start operating in 1999.

The 100-strong team of plasma physicists who make up the core of the laboratory has stayed basically intact over the past decade, even though the laboratory's total staff — and its budget — have fallen by half. But if the team is to remain together after this summer's further contraction, say fusion researchers, more will be necessary than just the construction of NSTX.

One need is for increased collaboration with the smaller experimental tokamaks at



End of an era? The tokamak at Princeton's Plasma Physics Laboratory is closing after 15 years.

General Atomics in San Diego, California, and at Massachusetts Institute of Technology (MIT), as well as with larger ones abroad — Joint European Torus (JET) in the United Kingdom and JT-60 in Japan. Princeton is seeking \$9 million next year to pay for such activities, including \$2 million for international collaboration.

Another need, according to Miklos Porkolab, director of MIT's fusion science programme, is for more machines — at least one medium-sized one. Princeton hopes to receive \$4 million next year to study three or four advanced tokamaks and alternative options, says Schmidt.

Perhaps above all, the laboratory will need a new director to forge a new direction. Will Happer, chair of the research board at Princeton University, which operates the laboratory, has been searching unsuccessfully since Ron Davidson, the current director, announced last autumn that he intended to depart.

"In this time of major changes in the US magnetic fusion programme, the job of the next director will be exceptionally challenging," says Happer. He points out that PPPL needs not just an exceptional scientist and manager but also "an articulate spokesman for plasma physics research who can help to unify not only the talented plasma physics community in Princeton, but the national community as well".

Given the magnitude of the challenges it faces, many observers feel that Princeton may find it hard to attract a top-flight external candidate to the laboratory — even after the existing management team has completed June's painful restructuring.

Schmidt says that the laboratory, two-thirds of which is structured around TFTR, will return to its previous organizational structure, with, following a long tradition at Princeton, "a combination of modestly sized experiments, broadening the science base".

Researchers at the laboratory are proud of the track record of the machine that shuts down next week. It has set world records for fusion power output, has burned deuterium and tritium together for the first time and has notched up a long list of milestones in plasma physics. "TFTR has been terrifically successful," says Davies.

NSTX will make use of beam drivers and power supplies already being used for TFTR. It will build on what Martin Peng, the programme director, describes as "encouraging" results from the Small Tight Aspect Ratio Tokamak (START), an experiment of similar geometry at Culham in the UK.

The new machine will be built quickly and cheaply, largely using existing components. "NSTX is a good project," says Bruno Coppi, a plasma physicist at MIT who is a critic of conventional tokamaks. "But it isn't highly original."

Coppi has long claimed that compact machines could achieve ignition at far lower cost than current tokamak designs, and suggests that PPPL can prosper by embracing innovative approaches to fusion power. If it does so, he says, "Princeton can use its intellectual strength to spring back".

But doubts about the future are widespread. "Princeton has one of the pre-eminent teams [of plasma physicists] in the world," says Alan Gibson, deputy director of JET. "Whether they can retain that position is open to question. It's going to be difficult."

Much of the US plasma physics community — including MIT's Porkolab — have trained or worked at PPPL and still fervently believe in the need for a strong laboratory there. "It is so easy to destroy, but will take decades to build up if you ever need it again," says Porkolab.

The task of the new director will therefore be daunting. As Porkolab puts it: "He has to find a viable concept beyond NSTX, stabilize the budget, and integrate the laboratory into the rest of the community." All this at a time when the shrinking of the US magnetic fusion budget combined with PPPL's efforts to retain a significant slice has led to friction between the laboratory and the rest of the US magnetic fusion community (see *Nature* 379, 387; 1996).

Now the laboratory is supposed to defend its own position, and provide leadership for the national programme at the same time. Its predicament makes fusion scientists hope for the best, while fearing the worst. "I'm sure it will not collapse," says Coppi. "There is so much knowledge there — it would be a disaster."

Colin Macilwain

UK parties differ on means, not ends

David Dickson

Many scientists are looking forward to a likely change of government, as the UK's election campaign begins in earnest. But Labour in power would face similar pressures and constraints to the Conservatives.

Over 30 years ago, when the then leader of the Labour party, Harold Wilson, promised to forge a new Britain in the "white heat of technological revolution", his words sent a tremor of anticipation through the research community. The election of a government explicitly committed to science and technology, it was felt, could bring only good news for researchers.

But in the harsher political and economic climate of the 1990s, similar rhetoric from the current Labour leader, Tony Blair, espousing the age of computer networking and gene sequencing, has aroused little enthusiasm in university laboratories.

Feeling those fiscal constraints

It is not for want of trying. Labour politicians frequently repeat a promise that they intend to place science "at the heart of government", along with education and training. Blair himself, in a key speech last year, promised to "end the bias against science in our culture", and to "strengthen the voice of science in government".

The problem is that, with self-imposed room for fiscal manoeuvre as tight as that imposed by the Conservatives, concrete promises have been kept to a minimum. Indeed, one of the characteristics of the six-week campaign set off by last week's announcement of a general election on 1 May is the lack of a significant difference in substance between the two major parties on issues relating to science.

Both, for example, are committed to the Technology Foresight programme (with the Labour party merely promising to try harder than the current government to ensure that its conclusions are implemented). Both favour selectivity and the creation of centres of research excellence as a way of maximizing the use of scarce research funds.

Both acknowledge that the market, if left to its own devices, is unlikely to guarantee an optimal level of research. And both want greater efforts in science education, and increased public understanding of science.

The differences between the two major parties tend to lie in means rather than ends. Perhaps the most controversial issue is how to introduce greater flexibility into government research laboratories. Conservative thinking still tends to favour privatization as

a solution. Labour talks more about rationalization, coordination and cost-effectiveness — although few details have appeared of what this would mean in practice.

In line with its main election strategy, the Conservative party is standing firmly on its record. Ian Taylor, the minister for science, points out that the reforms introduced in the science white paper of 1993, *Realizing our Potential* — which emphasized the needs of wealth creation and the quality of life — have successfully remoulded the organization of the science base to the demands of the 1990s. That conclusion was generally endorsed by a report issued last week by the all-party House of Commons Select Committee on Science and Technology.

Taylor points out that spending on research is growing in Britain at a time when it is decreasing in many of its competitors, and that Britain is also the most cost-effective producer of research among the advanced industrialized nations. He says that the wide acceptance of initiatives such as Technology Foresight indicates that the government has got its science policy about right — even if it cannot offer any large amount of extra money to researchers.

The boldest and most explicit proposals for change come from the smaller Liberal Democrat party, which has little chance of holding the reins of power but has committed itself to increasing the science budget by £150 million (US\$238 million) "because it has been squeezed for too long". It also pledges to spend 2 per cent of gross domestic product on R&D, and to set up an inquiry into mathematics education.

But all eyes are on the Labour party, currently riding high in the opinion polls and — bar some unexpected scandal — generally expected to win power for the first time in 18 years.

Part of Labour's strategy has been to refrain where possible from making concrete promises about what it will do once it is in power. This reflects voters' fears of possible tax increases, which are generally held to have lost the party the last election in 1992.

Despite constant criticism of the Conser-

vative government for cutting various areas of research spending, in particular that of government departments such as the Department of Trade and Industry, the Labour party has not made any commitment that it will reverse such trends.

"You don't always solve problems by throwing money at them," says Adam Ingram, the shadow science and technology minister and member of Parliament for East Kilbride.

Nor has Labour promised to reverse one of the most controversial decisions taken by the current Conservative government, namely the move of the Office of Science and Technology into the Department of Trade and Industry. The office had been set up in the Cabinet Office shortly after the last election (see panel, opposite page).

What Ingram does promise is that Labour will pursue an aggressive policy to ensure that the science base is linked to the needs of British industry through a range of initiatives from the enthusiastic promotion of Technology Foresight to a promise to examine tax incentives for corporate investment in research and development.

"Technology Foresight will be at the core of our strategy," says Ingram. "It is a way of focusing the minds of the university sector on what industry needs, and of building links between industry and the science base of this country."

Other influential voices in the Labour Party, including Anne Campbell, Labour MP for Cambridge and chair of the cross-party Parliamentary and Scientific Committee, are equally keen that the party, if it reaches government, should use its science policy to do more to encourage innovation — particularly in small companies — and build robust bridges between industry and universities.

Industry needs universities

The message has not gone down particularly well among many traditional Labour voters in the research community, who, while supporting its social priorities, would like to see the party protecting the academic world from the pressures of corporate influence, and tend to characterize Technology Foresight as the thin end of this particular wedge.

But, like most of the leaders of 'new' Labour, keen to stress their commitment to industrial innovation, Ingram, who trained as a computer programmer and comes from a trade union background — he was an official of the local government union NALGO for ten years — has little time for such sensibilities. "People think that you can continue with the old formula and the old structures, but they are not there any more," he says. Universities will have to learn to pay their way. "If our universities do not meet the needs of industry, there will not be jobs for those coming out of the university sector."



Taylor: points to policy record.

Nor is the Labour party, despite its egalitarian social philosophy, being squeamish about calling for a greater concentration on high performers when it comes to the distribution of funding for research. A director of one of Britain's leading pharmaceutical companies says that his biggest problem with a Labour government would be if it started "levelling downwards" in allocating research funds. But this seems unlikely, as the Labour party now espouses concentration and selectivity with, if anything, even greater fervour than the Conservatives.

"We must focus on the best," says Ian Gibson, dean of biological sciences at the University of East Anglia, who is standing as Labour candidate for the marginal seat of Norwich (North). Gibson, who raised the score of his department in the government's research assessment exercise from 3 in 1992 to a 5 in 1996, and if elected will join a steadily growing band of scientifically-trained Members of Parliament, says that a commit-

ment to good research is essential. "You may have to be really hard about it, and be prepared to say that the best is the best".

Perhaps wisely, the issue where attitudes to selectivity will have the greatest impact — namely the future organization of universities — has been 'kicked into touch' as far as the election is concerned by the government's decision last year, endorsed by Labour, to pass it to an inquiry headed by the former civil servant Sir Ron Dearing.

Opening and opportunities

It will be less easy for Ingram, or whoever else is awarded the science portfolio, to avoid pressures on the science base to cut costs and increase cost-effectiveness. The easy political points have been made by expressions of outrage at the government's recently abandoned efforts to implement the wide-scale privatization of government research laboratories (see *Nature* 385, 473; 1997).

But pressures will remain to eliminate

research in government-funded institutes that is considered insufficiently productive — a move that is likely to put a Labour government in conflict with the unions, as redundancies could well be involved.

Ironically, therefore, even those scientists who support the Labour party accept that they are unlikely to be substantially better off under a Labour government. But there is still optimism that the party may be able to fill perceived gaps in UK science policy — such as the lack of strong coordination between government departments, or a robust strategy to defend those on short-term contracts.

And even those who remain disappointed at the lack of manifesto commitments remain confident that there will be an opportunity to influence policy once the election is over. "All bets are off until then," says one active member of the support group Scientists for Labour. "After that, our hope is that, if we shout loud enough, they will listen."

Little change in the politics of science — but closer links with industry?

[LONDON] One of several areas in which the expected election of a Labour government in Britain is likely to have minimal impact on British science is the allocation of political responsibility for the science base.

Two years ago, the Conservative government shifted this responsibility from the Cabinet Office, symbolically close to the prime minister, to the Department of Trade and Industry (DTI). The move produced howls of outrage at the apparent demotion of science in the political pecking-order (see *Nature* 376, 103; 1995).

Many scientists who support the Labour party would like to see the decision reversed, or indeed for science to be given its own government department, arguing that this is necessary to meet the party's promise to put science "at the heart of government thinking".

Robin Walters, for example, a researcher at the University of Sheffield and secretary of the 300-strong group Scientists for Labour, says one of the group's priorities is to ensure that science is given full cabinet status.

A similar demand has been made by the lobby group Save British Science. There has been some response in political circles, notably from the Liberal Democrats, who are committed

to moving the Office of Science and Technology (OST) back to the Cabinet Office.

But Labour has wavered. At times in the past, it has been in favour of a separate minister for science. This time, however, while criticizing the haste and lack of consultation about the move to the DTI, it has no plans to change it.



Ingram: 'working with industry'.

According to Adam Ingram, shadow minister for science and technology, part of the reason is that, after four years of upheaval which started with the publication of the science white paper in 1993, the party has no desire to impose further disruption on the scientific community.

An equally powerful reason, however, lies in the fact that Labour is committed to forging even tighter links between the science base and the private sector than the Conservatives. This commitment is summed up in Ingram's words that a Labour government will work "with and on behalf of industry". From this perspective, placing OST within the DTI makes sense.

As with Labour's broader policies, this strategy has met a

mixed response. Many scientists fear that leaving the OST where it is risks increasing pressure on research councils to tailor their activities to an industry-led agenda.

But Ingram is unapologetic, arguing that "people should not sit in their ivory towers and be frightened of industry, because if industry does not work effectively there is no future for a strong science base in this country".

Where Labour does promise change is in the coordination of research policies across government departments. It is widely agreed that the reforms promised in the white paper have made relatively little impact in this area of activity.

In principle, such coordination is the responsibility of the head of the OST, who doubles as the chief scientific adviser to the government. In practice, the OST's relative lack of leverage over the research activities of high-spending departments has only been exacerbated by the move to DTI.

Anne Campbell, Member of Parliament for Cambridge and an influential voice within the party, says: "I would like to see the chief scientific adviser's position strengthened, so that he really is a 'Mr Science' in government, with more authority over the other scientific advisers, and the

ability to ensure that the government's overall strategy for science is better followed."

Ingram concurs, arguing that such an individual would, in practice, be in a better position to achieve such coordination than a cabinet minister because he or she would be able to use arguments about the need for high-quality science and science advice to avoid the "internecine warfare" between government ministers.

As a step in this direction, Labour has already promised to appoint a separate chief scientist at the DTI, a post abolished by the Conservatives in the 1995 reshuffle, and now occupied in principle by the head of the OST, who doubles it with his role as adviser to the prime minister.

Further changes will have to await the outcome of what Ingram describes as a "period of quiet reflection" which the Labour party has promised will follow the election if it wins — and which some feel has provided an excuse for its reluctance to offer firm plans before the election.

Meanwhile Ingram says that there are no plans to replace Sir Robert May, the chief scientist, or Sir John Cadogan, director general of research councils. Both are civil servants and entitled to remain in their posts after the election.

D. D.

Nuclear waste body cancels contracts after plans rejected

[LONDON] Nirex, the company charged with disposing of waste from Britain's nuclear power stations, has cancelled contracts for an underground 'laboratory' to investigate the feasibility of storing waste beneath Longlands Farm in Sellafield, Cumbria.

The news follows last week's decision by the government to dismiss Nirex's application to build the laboratory, partly because of scientific uncertainty about the characteristics of the proposed site. Contract staff working on the laboratory have been asked "to return to their parent companies". Nirex's chief executive, Michael Folger, says the government's decision may also have implications for the company's own staff.

MAFF under fire on cloning funds

[LONDON] An all-party group of British politicians has urged the Ministry of Agriculture, Food and Fisheries (MAFF) to reconsider a decision, which it describes as "cavalier and blinkered", to withdraw funding for the research that led to the first successful cloning of an adult sheep at the Roslin Institute in Scotland.

The House of Common select committee on science and technology has also suggested that British legislation should be altered to close a loophole in a law passed in 1990 that had been intended to ban all cloning of human beings, but whose language left ambiguous whether it applied to the techniques developed at Roslin (see *Nature* 386, 98; 1997).

Company heads plead guilty on AIDS charged

[TOKYO] In a surprise move, three former presidents of Midori Juji (Green Cross Corporation), Japan's leading blood product manufacturer, pleaded guilty on Monday (24 March) to charges of professional negligence resulting in the death of a liver patient from AIDS, the first admission of guilt in the recently opened criminal trials in Japan's HIV blood scandal.

The three former presidents were expected to plead innocent, but apparently concluded that the prosecutor's case was too strong. The admissions of guilt are expected to strengthen the hand of prosecutors in other criminal cases filed against Takeshi Abe, the head of the government's AIDS study group in 1983, and Akihito Matsumura, a former government official. Both claimed innocence at their first court hearings two weeks ago.

Cassini mission faces new safety questions

[WASHINGTON] NASA plans to update its Environmental Impact Statement for the Cassini mission to Saturn, reopening to public scrutiny questions about the safety of onboard plutonium batteries. A 1995 impact statement showed that the batteries — used to provide power at great distances from the Sun — pose a negligible risk in the event of a launch accident or spacecraft re-entry. But new information has since come to light, according to Mark Dahl of the agency's space science office. Dahl will not say what the new information is, but says "the [overall] bottom-line risk is about the same". A draft of the 'supplemental' impact statement will be available for public comment in April. Anti-nuclear groups plan to protest against the Cassini launch, set for October.

Engineer to head Italian agency

[MUNICH] Lucio Bianco, a 55-year-old engineer, has been nominated as the next president of the CNR, Italy's major basic research organization. He takes over from Enrico Garaci next month, once his appointment has been approved by parliament, which is a formality. He will preside over major changes intended to



Malcolm is imp
by the new Pre
Protease be
it's a GST
fusion prot

"Not being a protein chemist, I just want to clone the gene, express it, isolate the protein and move on," says Malcolm Zellars, who's working on his post-doc at Tufts University Medical School in Boston, Massachusetts, USA.

Not the fault of Confucius

Sir — P.-L. Chau and W. Y. Chau (*Nature* 385, 110; 1997) dissent from a leading article (*Nature* 384, 197; 1996) apparently linking the lack of creativity in Korea to a culture deeply influenced by Confucian thought. But who said Confucian-influenced Asians have to be 'non-creative' or even Confucian? We in Taiwan are "proud" to report that, in response to the sheep-cloning experiment in Edinburgh (*Nature* 385, 810–813; 1997), some of our government officials and faculty members in leading universities here openly declared that Taiwan scientists had actually either "preconceived" the idea or "done it all". The following are some excerpts of what has been said.

"The technique is very simple. In fact, we cloned two mice several years ago." That was said by a professor at the School of Medicine, National Taiwan University (NTU). When asked how it was done, he refused to divulge any details on the astonishing ground that these details belong to a "top-secret X-file" (*Central Daily News* 5 March 1997, page 8).

"It is unfair to our country's scientists!", complained the deputy director of the Animal Industry Department, Council of Agriculture, Executive Yuan of Taiwan. "Why is everyone only interested in the British sheep-cloning work?" He then asserted that several institutes in Taiwan, for example the Department of Animal Sciences of NTU, the Pig Research Institute and the Taiwan Livestock Research Institute, had done similar or identical studies, and lamented that no one seemed to have noticed. To top it all, he began with praise of one "best" professor from the NTU, who apparently did his graduate work in the United Kingdom, and followed it with a startling, yet wistful, speculation: "Had this professor chosen to stay in England, it is highly probable that he would be the one basking in the glorious limelight" (*The Independence Morning Post* 5 March 1997, page 7).

So far, these outrageous statements have drawn no challenge from other scientists. To scientific communities in Western countries, this baseless bravado and the muted response

from the science establishment here can be easily dismissed. But to us Asians, especially those of us who do scientific research here in Taiwan, it is a far more serious problem. For, first of all, as had been pointed out in the Chau letter, Confucius taught us the virtues of modesty and intellectual honesty of which such bravado is the antithesis. This behaviour, in fact, might be called "Ah-Q-ism" as it has all the characteristics of the anti-hero in Lu Xun's novel *The Story of Ah-Q*, an absurd parable about the Chinese populace around the 1920s. More disturbingly, the silence of the Taiwan scientific community reflects a deplorable reality. Subconsciously, Ah-Q-ism has not yet been cleansed from our national psyche. What else could explain a malady that drives them to crave for or tolerate such self-delusion? Until we have the will and the determination to do such cleansing, it is premature and perhaps even absurd to blame Confucius for our collective woes in science.

Klim King, Young-Sun Lin, Hong N. Yeung
Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, Republic of China
e-mail: bmkxk@alpha.sinica.edu.tw
bmynslin@ccvax.sinica.edu.tw
bmhny@ccvax.sinica.edu.tw

No lack of interest

Sir — I was both surprised and shocked by your recent comments about Edith Cresson, the European commissioner for research, education and training (*Nature* 385, 661; 1997). As I have myself held this post, I continue to follow the development of research policy. I have also, as chairman of a European Union panel, recently evaluated the current Framework programme and submitted a report to the commission and to the European Parliament. I have had meetings with Cresson to discuss these issues at some length, and have also had repeated contact with various key figures in the scientific community and industry directly interested in research policy and the way in which it is conducted.

I can, of course, speak only for myself, but I am surprised that you feel that the commissioner demonstrates a "shameful lack of interest in science". Cresson is moving to strengthen research policy and to increase its effectiveness and achieve better support for European competitiveness.

I should like to mention just a few of the initiatives I am aware of. Cresson set up, with her colleagues Neil Kinnock and Martin Bangemann, task forces intended to bring research workers, industrialists and users closer together in sectors such as manufacturing and cars, or in areas essential for the environment and health, such as water management, multimodal transport and research on viral diseases. Cresson also

reduced bureaucracy by creating a simplified standard research contract.

As a member of the European Round Table of Industrialists, we have drawn the attention of the commission to the importance of education, and our cooperation has led to a very important initiative taken by Cresson about "learning in the information society".

Cresson has also initiated a process of consultation on an unprecedented scale as regards the future of community research policy, as part of the preparations for drawing up the commission's proposals for the fifth Framework programme. Since last summer, three successive orientation documents have been submitted for discussion to the institutions and authorities concerned, political leaders, elected representatives, research workers, industrialists and the press.

As president of the evaluation panel of the Framework programme, I can testify that Cresson's proposals do not aim to "repackage" the current Framework programme.

To blame Cresson — or indeed the commission as a whole — for the sluggishness of the Framework programme's adoption procedure is unfair, as it is the treaty on European Union, signed by the member states, that set out this procedure, as we have pointed out in our evaluation report.

I feel the criticism directed towards Cresson may be caused by the unavoidable

resistance and opposition that any strong personality trying to introduce change always runs into.

Etienne Davignon

*Société Générale de Belgique,
Rue Royale 308,
B-1000 Brussels, Belgium*

Unclothed emperor

Sir — The postmodernist ideas about the nature of scientific knowledge described in a recent News story (*Nature* 385, 381; 1997) are not new and have already been overtaken by the history of science. I was educated in Czechoslovakia during the Lysenko era. Empirical genetics was considered an ideological construct of the bourgeoisie (the term then used was 'superstructure'). The "bankrupt capitalistic genetics of Mendel and Morgan" was contrasted with the "victorious proletarian genetics of Lysenko", and Mendel's experimental garden changed into a volleyball court. How the story ended, everybody knows. The emperor of postmodernist sociology of science not only has no clothes, he is not even new and original in his nakedness.

Viktor Holoubek

*Department of Human Biological
Chemistry and Genetics,
University of Texas Medical Branch,
Galveston, Texas 77555, USA*
e-mail: vholoubek@mspo1.med.utmb.edu

Misplaced complacency about energy resources

Sir — In his letter “Flawed reasoning about oil and gas”, Maurice B. Dusseault (*Nature* **386**, 12; 1997) criticizes the review by Peter Kassler of a book by Marcello Colitti and Claudio Simeoni entitled *Perspectives of Oil and Gas: The Road to Independence* (*Nature* **384**, 528; 1996). Both the review and the letter are really having a go at writers who face the finiteness of fossil fuels and seek to stir up some discussion about how the human race will organize the economics under the threat of scarcity.

Industries and big centres of population are built around indigenous resources, and they have ultimately to look elsewhere, usually in unindustrialized countries, for resources so that they can continue their enterprise. A pattern of life is thus developed that depends very fundamentally on a continuing supply of energy.

Neither writer recognizes that global pollution has become a pressing problem. Kassler is not bothered by the difficulty of the continued supply of fuel, although he has a slight worry because oil supply has become a “political football”. He relies on a vigorous free market “to keep my children’s feet warm in their old age”.

Dusseault pretends, typically of an economist, that there is no problem because the Canadian heavy oil deposits are “sufficient to meet current US and Canadian consumption combined for more than a hundred years”. Canadian “political and military disruptions are modest, compared to recent Middle East events”. The struggle for political power over oil supplies and the problems of climate change and local pollution, due to using the residual difficult mineral fuels, are all seen as technologically soluble now and within the foreseeable future.

Environmentalists and others, who see humanity’s future as being at least as extensive as our recorded historical past, do not see the advances achieved in the past two-and-a-half centuries being simply put into reverse. Even if the world population is stabilized as soon as 2050 at 10 billion, it will not be living in the right places for the development of the shales. And even if we have all learned to live on orimulsion and try to cultivate the world like a garden, it will lack its forests and we shall be the enemies of all other species — except possibly for the flocks of clones we enslave to serve our crude purposes (in a vigorous free market — or will it be a battlefield?).

R. S. Scorer

*Department of Mathematics,
Imperial College,
London SW7 2BZ, UK*

Sir — In recent months, *Nature* has brought us two highly imaginative alternatives to conventional fossil fuel. The first, where petrol is produced by “adding leaves and bark extracts from a native herb to tap water,” has tragically turned out to be a hoax (*Nature* **383**, 112 & **384**, 106; 1996). Let’s hope that Dusseault’s suggestion fares better.

Pointing out that “as conventional oil deposits are gradually depleted, prices will rise,” thereby making “non-conventional resources economically accessible”, Dusseault states that “one may even extract carbon (from atmospheric carbon dioxide or limestone) and hydrogen (from solar-powered electrolysis of water) and assemble synthetic gasolines from the basic elements”. He adds: “The only realistic limitation to this process is price.”

While I salute Dusseault for his bold attempt to liberate us from the constraining fetters of thermodynamics, I think he has overlooked an even simpler solution to the problem. All cars come equipped with a battery and an electric starter-motor. So why not dispense with petrol altogether and let the starter-motors do the work?

Douglas Axe

*Department of Chemistry,
University of Cambridge,
Cambridge CB2 1EW, UK*

Habitable moons

Sir — The recent discussion of habitable moons by Williams *et al.* and by C. Chyba (*Nature* **385**, 234–235 & 201; 1997) is prefigured in some science fiction, with implications for the extent of the habitable zones around stars.

In 1975, Poul Anderson wrote a description of such a habitable moon to form the background for a set of stories written as part of a science fiction seminar at the University of California, Los Angeles. (The essay and the stories are collected in *Medea: Harlan’s World*, ed. H. Ellison, Bantam, 1985.) In this he pointed out that, as large gas giants are reasonably hot bodies themselves, they could warm the moons orbiting them. Thus a moon orbiting a superjovian planet outside the normally accepted habitable zone might be able to support liquid water, thanks to the added heat flux from its primary. The habitable zone for moons could thus be larger than that for solo planets.

In a separate essay undertaken for a similar exercise in collaborative world building (Murasaki, ed. R. Silverberg, Bantam, 1992), Anderson points out that the habitable zones around small red dwarfs will be very close to the star and that small planets within this zone may well be tidally locked. Having one permanently heated side and one cold side may render the ‘habitability’ of such planets somewhat

theoretical. Anderson’s solution to this problem was a pair of rocky planets in orbit around each other. An alternative scenario for true habitability in such a close orbit would be a rocky moon. It could be tidally locked to its own primary, rather than to the star, and thus could have a less problematic day/night cycle, perhaps of the order of ten days rather than 100. It would also benefit some of the time from night-side warming by its primary.

Oliver Morton

*3 Hoskins Street,
London SE10 9NZ, UK
e-mail: omorton@wired.co.uk*

Malicious referees

Sir — Simon Wain-Hobson is by no means the first to express dissatisfaction with the refereeing of papers, and won’t be the last (*Nature* **385**, 384; 1997). I second his proposal for reviews to be of a sufficient length — say 10 per cent of the paper’s length for a rejection, 5 per cent otherwise?

There is another change that would not only improve the quality of reviews and consequent decisions but also make the editorial decision-making process more efficient. The idea is that editors should not examine referees’ reports initially but should send them immediately to authors, and then if, and only if, the author feels he or she has a credible defence against the reviews will she or he then submit his or her response to the editor for consideration. Reviewers would then know that they could not get away with sloppy or malicious reviews. And editors would be saved the work of trying to weed out poor papers that can slip through the less controlled review process that is the current norm. This could go some way towards reducing the massive volume of poor-quality papers being published.

Meanwhile, authors could copy the trick I once used of compiling a collection of previous reviews of the paper along with my rebuttals, and sending that with the submitted manuscript to the next journal. It worked for me. The paper involved was described by a reviewer for *Personality and Individual Differences* as well-written, well-argued and well-documented, whereas a *British Journal of Psychiatry* reviewer reckoned it was of lowest grade in all three respects!

It is difficult to see how this gross discrepancy can be accounted for except in terms of malice on the part of the latter reviewer, such as I encountered with many previous reviews.

R. P. Clarke

*Abnormal Psychology Research Institute,
9 Augusta Road East,
Moseley, Birmingham B13 8AJ, UK
e-mail: rclarke@fineart.demon.co.uk*

Worldwide ageing programme needed

Sir — I read with interest the News story about the issues raised by Japan's demography, showing that a dramatic increase in population age is being accompanied by a decrease in the number of financially productive individuals (*Nature* **385**, 379; 1997).

The extrapolations that can be made from Japan to other countries and possibly to the entire world are striking. Scientists are actively searching for ways to increase the human lifespan and at the same time to solve the medical problems of old age, such as degenerative disease and dementia. But old people must also have a socially and economically acceptable quality of life. Several questions need answering. Is it reasonable that scientists should try to increase the length of human life without a parallel change in the mechanisms by which society treats the aged? Would it be ethical to cease these efforts? Can state health insurance systems insure everybody against the as yet inevitable diseases associated with old age, without violating the very principle of 'insurance'? Can an old person still be active and productive for society?

The answers to these questions cannot come from single disciplines. Some international organization should take responsibility for coordinating all aspects of ageing research, cultural, social and biological. Unesco has shown itself effective and successful in a great number of worldwide projects carried out by its three major sectors — education, science and culture. Could Unesco unite efforts in a global project of providing, for the year 2000, a convincing model to tackle the population ageing problem? Or do we rather need a new organization, a United Nations Programme for Aged Persons, perhaps?

Angelo Azzi

Department of Biochemistry,
University of Bern,
Bern, Switzerland
e-mail: azzi@mci.unibe.ch

Data available

Sir — Our laboratories are currently determining the complete nucleotide sequences of the genomes of the obligate human pathogens *Streptococcus pyogenes* and *Neisseria gonorrhoeae*, sponsored by the National Institute of Allergy and Infectious Diseases (NIAID) of the US National Institutes of Health (NIH). As we have been committed from the beginning of these projects to an early data-release

policy for the benefit of the scientific community, we wish to notify the community of the availability of preliminary nucleotide sequence databases for these continuing genome sequencing projects. Contrary to recent assertions in *Science* (275, 777; 7 February 1997), these data have been freely available on the Internet since mid-September 1996 and may be obtained at <http://www.genome.ou.edu>.

The data collected for both organisms have been compiled into Blast databases that are searchable from our Web site; alternatively, the data can be downloaded via FTP from this site. These databases represent raw, assembled data that are not currently proof-read, provided in a series of contiguous segments (contigs), which will decrease in number as the gaps between the contigs close. All data should be considered preliminary until gaps are closed and proof-reading and annotation are complete.

Investigators using any of this information are asked to acknowledge these microbial pathogen genome projects, sponsored by NIAID/NIH. Further information can be obtained from J.J.F. (*S. pyogenes*), or D.W.D. (*N. gonorrhoeae*) or from B.A.R. (both organisms) at the addresses below.

Joseph J. Ferretti

David W. Dyer

Department of Microbiology and Immunology,
Oklahoma University Health Sciences Center,
Oklahoma City, Oklahoma 73104, USA

Bruce A. Roe

Department of Chemistry and Biochemistry,
University of Oklahoma,
Norman, Oklahoma 73019, USA

Worthless ranking

Sir — Steve Jones's analogy between Gogol's *The Government Inspector* and the Research Assessment Exercise (*Nature* **385**, 311; 1997) was just beautiful. Perhaps the point is that science is done by individuals, not by departments. To rank science by department is about as clever as ranking journals by their impact factor; neither procedure tells one anything very interesting (except to the bureaucratic mind). The former is a crude average over a group of very different people which tells one little about the science done by each of them. The latter is a method of ranking the commercial success of journal editors, not a method for assessing the worth of individual scientists. (My own department was 5-star, so I have no axe to grind.)

David Colquhoun

Department of Pharmacology,
University College London,
Gower Street,
London WC1E 6BT, UK

Marijuana as a medicine

Sir — Your recent News story on marijuana (*Nature* **385**, 756; 1997) was factual and objective but, as in most of the other media coverage of the subject of smoked cannabis versus oral Δ -9-tetrahydrocannabinol (THC) (Marinol), an important point has been omitted.

The cannabis plant is a complex mixture of substances that include at least 60 different cannabinoids, some of which have been shown to have pharmacological activities of their own. For example, one such component, cannabidiol (CBD) can affect the metabolism of THC and, therefore, in some individuals may improve the anti-emetic properties of the latter. Interestingly, cannabidiol is usually more abundant in most samples of marijuana.

Given such facts, it seems rather unenlightened to dismiss the reports from cancer and AIDS patients who claim greater benefit from smoked marijuana than from oral THC as simply a ploy to facilitate government approval for medical use.

Sumner Burstein

Department of Biochemistry and Molecular Biology,
University of Massachusetts Medical School,
Worcester, Massachusetts 01655, USA

Hantavirus in Argentina

Sir — You say in a recent News article about staff and budget cuts in Argentinian health laboratories that there are 20 new cases of hantavirus infection each week (*Nature* **385**, 376; 1997).

We are working at a private hospital in the city of Bariloche, where most of the cases have been treated. Up to 15 December 1996, 77 cases of HPS (hantavirus pulmonary syndrome) had been reported and confirmed in Argentina. Of these, 37 people died and the rest recovered (D. Enria *et al. Medicina* **56**, 709–711; 1996). During 1996, a total of 22 cases were notified in southern Argentina. But no new cases have been reported this year.

The economy of the whole lake district on the eastern slopes of the southern Andes and centred on Bariloche is based on tourism. The region has suffered its worst summer season ever because of misleading media accounts about the extent and severity of the hantavirus outbreak.

Nevertheless, we are deeply concerned about the staff and budget cuts to the National Institute for Microbiology, an institution that has played an important role in handling the hantavirus outbreak.

Carlos M. Barclay, Gabriela Rubinstein

Sanatorio San Carlos,
Av. E. Bustillo Km. 1,
8400 Bariloche, Argentina

Language that dare not speak its name

Proposals by a school board in California to recognize the dialect used by most of its pupils unleashed a ferocious media attack. Why did the press get things so wrong, and why were the proposals so virulently ridiculed?

Geoffrey K. Pullum

Every columnist in the English-speaking press seems to have had a say about last December's resolution by the governing board of the Oakland Unified School District in California. The board's proposal was to recognize the native tongue of most of its pupils as a language, which triggered an astonishingly ferocious media attack. But little was heard in the press of the factual matters about language and learning that underlie the controversy.

The *New York Times*¹ (echoed by the *Economist*²) reported that the board "declared that black slang is a distinct language". The board did nothing of the kind. Slang, in any language, consists of a finite list of words or idiomatic phrases, highly vivid and informal, in the most casual stratum of its lexicon. The resolution neither stated nor implied an interest in the ephemera of street slang. It was about a perfectly ordinary variety of English spoken by a large and diverse population of Americans of African descent, by no means all of whom are slang-users.

The board mentioned several names for this language, one of which came to dominate the headlines: "Ebonics", introduced in 1975 by the African-American scholar Robert L. Williams³. This term's blend of 'ebony' and 'phonics' suggests confusion: phonics is a literacy teaching method, not a language, and racial characteristics such as ebony skin do not correlate with linguistic boundaries⁴. The misbegotten name was a public-relations disaster, lending itself to endless puns and perversions — such as rhetorical questions about whether Jewish speakers of English would declare a new language called "Hebonics".

Much ink was wasted on the question of whether the speech of African-Americans should be classified as a language separate from English. Defining a language is not a trivial problem, especially as political considerations are often involved. For example, what was once a single language called Serbo-Croat is being split into three separate languages (Serbian, Croatian and Bosnian) to serve political constituencies. And although one standard source recognizes more than 15 separate Romance languages in Italy, another recognizes just three.

But essentially all linguists agree that what the Oakland board was dealing with is a dialect of English. I will refer to it as African-American English (AAE) — though of course it is not spoken by all African-Americans; many speak only standard English.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Linguistic ghetto: to dismiss African-American English as 'bad' English reflects a lack of understanding of the nature of language.

The general public seems unaware that AAE is regular, stable and governed by rules of grammar and pronunciation that are as consistent as those of any other spoken language. It differs strikingly from the standard dialect, but there is no more reason for calling it bad standard English than there is for dismissing western dialects of English as bad eastern speech, or the reverse.

Yet AAE is constantly described as if it were English with mistakes and omissions. For example, it is alleged that the copula (*am, are, is, and so on*) is 'omitted' in AAE — carelessly left out, critics seem to imply — or incorrectly used in the uninflected form *be*. This is a jumble of falsehoods. The rules regarding the AAE copula are quite detailed and specific. If it bears stress, as it always does at the end of a phrase, it must be pronounced. So it is obligatory in *Couldn't nobody say what colour he is*, meaning 'Nobody could say what colour he is' (examples in this paragraph are from the utterances of Larry, a speaker studied by William Labov⁵). If the copula is negated, it cannot be omitted; it is pronounced *ain't*, as in *You ain't goin' to no heaven* ('You aren't going to any heaven'). There is a special habitual aspect that standard English lacks, and for this *be* is used and must never be omitted: *they be sayin'* means 'they habitually say'. The copula is also obligatory when it is in the past tense. Only when the AAE copula is not marking habitual aspect, not negated, not stressed and in the present tense can it be unpronounced (*he good* means 'he's good'). Russian, Hungarian and other languages have a very similar rule (see box overleaf).

Something similar is true of the multiple

marking of negation, misleadingly termed 'double negation', and treated as an illogicality. This feature is also found in Romance languages. There is a direct analogy between AAE's *Ain't nobody called* and the equivalent Italian *Non ha telefonata nessuno*, literally 'not has telephoned no one'. Both mean 'No one has telephoned'. It is a rule in both languages that under certain conditions indefinite words with meanings like 'someone' or 'anyone' must be replaced by their negative counterparts when they occur in a negated clause. No grammatical or logical mistakes are involved; multiple negation marking is a grammatical requirement like number or gender agreement.

Having known such facts for decades, linguists were dismayed to see writers of all persuasions (several of them black) falling over each other to publish angry and offensive attacks on AAE in the wake of the Oakland board's announcement. Confusing lexicon with syntax, accent with dialect, difference with deficiency, and grammar with morality, commentators clarified little except the deep hostility and contempt whites feel for the way blacks speak ("the patois of America's meanest streets", columnist George Will⁶ called it, as if AAE could only be spoken in slums), and the deep shame felt by Americans of African descent for speaking that way (a *Los Angeles Times* column⁷ by Eldridge Cleaver, a former Black Panther party official, compared the official acknowledgement of AAE with condoning cannibalism).

Vying with each other to express their fury at AAE, columnists, both black and white, ignored the genuine issues of educational policy that had motivated the Oakland board. One persistent confusion was perhaps stimulated by an ambiguity in English: 'instruction in French' has two meanings: instruction on how to speak French; and instruction given via the medium of French. The press never figured out the difference. The Oakland board members talked about using AAE to deliver instruction; but this was discussed in editorials as if the proposal had been to make AAE a school subject. The board never suggested adding AAE to the curriculum. Its plan was to acknowledge that many Oakland schoolchildren speak AAE, to alert teachers to the implications this might have, and to contrast AAE with standard English in language classes, but not to hold classes on how to speak AAE. (The whole point, after all, is that no such classes are necessary, since the children arrive speaking it.) The need to direct children toward a command of standard English was a paramount concern.

Oakland faces an issue similar to the dilemma of, for example, a Norwegian school with many Swedish-speaking students. Swedish and Norwegian are similar enough that both could be considered dialects of a single language. They are mutually intelligible for adults, but different enough to cause some learning difficulties for a monolingual Swedish child confronting a Norwegian-speaking teacher for the first time. So would it be better to start the Swedish children off with a teacher who acknowledges the Swedes' linguistic background and explains things in Swedish where necessary? Or should one trust to the 'total immersion' philosophy and plunge the students into a Norwegian-only environment from day one?

These are empirical questions. And for some time there has been empirical evidence that alternatives to total immersion may work better. Teaching children to read first in their nonstandard Swedish dialect and then transitioning them to standard Swedish speeds and improves the acquisition of reading skills⁸. African-American college students in Chicago who received instruction concerning the contrasts between AAE and standard English grammar showed improved standard English writing skills as compared to a control group⁹. And in Oakland itself it was found nearly 25 years ago that teachers who condemned AAE pronunciations and interpreted them as reading errors got the worst results in teaching black children to read, while teachers who used AAE creatively in class got the best results¹⁰. These and other results in the education literature (cited by Stanford AAE researcher John Rickford; see <http://www-leland.stanford.edu/~rickford/ebonics>) suggest that the Oakland school board's policy decision had some clear motivation and scientific support.

The board's reward, however, was a month of unrelenting ridicule, needling and abuse from politicians, poets and pundits in editorials, features, talk shows, news programmes, speeches and policy statements (the US Department of Education pointedly announced that no federal bilingual education funds would be spent on AAE).

The horror with which Americans react to the idea of using AAE in the classroom reveals a lot about the prejudice still targeted on America's black citizens, whose variety of English is decried as if it were some repellent disease (the *Economist* actually entitled its article "The Ebonics virus"). Educational conservatives often deny that prejudice is involved, dismissing linguists' objective attitude toward nonstandard dialects as if it were just left-wing propaganda. But it is not. Even conservative linguists acknowledge the facts mentioned above. When the Linguistic Society of America voted in January on a resolution in support of the Oakland school board, the vote was unanimous.

Expression of the copula in three types of natural language

Zero copula languages

(present-tense copula unexpressed under certain grammatical conditions)

African-American English (Indo-European)	● unexpressed in nonemphatic nonhabitual present affirmative when not stranded
Arabic (Afroasiatic)	● unexpressed in nonemphatic present affirmative
Hungarian (Uralic)	● unexpressed in third-person present indicative affirmative
Russian (Indo-European)	● unexpressed in nonemphatic nonexistential present indicative affirmative
Swahili (Niger-Congo)	● unexpressed in nonemphatic present affirmative

Variable copula languages

(present-tense copula amalgamates with adjacent word)

Colloquial standard English (Indo-European)	● reduced to single consonant in nonemphatic affirmative present tense; independent verb elsewhere
Turkish (Altaic)	● reduced to suffix in present tense; independent verb elsewhere

Overt copula languages

(copula always overtly expressed as an independent word)

French, German, Italian (Indo-European), Abkhaz (Northwest Caucasian), Japanese (Japanese-Ryukyuan), and so on

African-American English (AAE) is dismissed as 'bad' English because the speakers sometimes 'omit' the copula from predicative clauses. But this is not random carelessness. AAE exemplifies a widely attested grammatical pattern: the copula is always expressed if emphasized, negated, in the past tense, marking the habitual aspect, or stranded (at the end of a clause, as in *I don't know who he is*). A similar pattern is found in many languages in the Afroasiatic, Indo-European, Niger-Congo and Uralic families. (These families are *not* known to be related, so the similarities are typological rather than traceable to common origins.) Standard English shows a different pattern, rather like Turkish, where the present-tense copula is not completely omitted but is radically reduced in pronunciation (in Turkish, to a suffix, and in English, to a single consonant, as in *he's tall*). In a third group of languages the copula is always overtly expressed.

Linguists agree that standard English has prestige and AAE does not. They merely note that this is not because of the grammar of AAE. Grammar does not underpin social distinctions or justify morally tinged condemnation of nonstandard dialects. AAE has nothing inherently wrong with it as a language; it is an accident of history that it is not the standard language of the United States. Standard English lacks multiple negation marking, but has syllable-final consonant clusters and interdental fricative consonants. In Italy the facts are the reverse: standard Italian lacks syllable-final consonant clusters and interdental fricative consonants but exhibits multiple negation marking, exactly as in AAE.

But the feeding-frenzy in the press over the Oakland resolution alarmed enough parents that, fearing linguistic ghettoization of their children, they pressed the school board to reconsider. And so on 15 January the board revised its statement, dropping the reference to AAE as the "primary language" of many of its schoolchildren, and the idea of using that language as a medium of instruction. Yet it retained some misguided references to alleged African origins of AAE (the dialect did not originate in Africa, and shows few clear signs of any influence the original African languages of slaves may have had on its early history). Carefully saving some Afrocentric bathwater, the board publicly threw out the baby of a sensible change in language policy.

However, the Oakland school system is apparently still intent on using AAE to assist the learning of standard English, despite the

board's apparent backdown. This should not be surprising. Several other school districts in California and elsewhere have long operated such programmes. And of course teachers have some discretion, and many Oakland teachers are African-Americans who speak AAE natively; it would be difficult to police classrooms tightly enough to prevent them from using their native tongue when speaking all day to children who share it.

African-American English will continue to be heard in Oakland's classrooms, just as German and Chinese accents will continue to be heard in the physics lecture halls at the University of California campus just up the road in Berkeley. In practice it will matter very little whether official policies (or newspaper columnists) laud or deplore the linguistic diversity of which the United States sometimes seems so ashamed. □

Geoffrey K. Pullum is a professor of linguistics at Stevenson College, University of California, Santa Cruz, California 95064, USA (e-mail: pullum@cats.ucsc.edu).

1. *New York Times*, 24 December 1996.
2. *Economist* p. 40, 4–10 January 1997.
3. Williams, R. L. *Ebonics: The True Language of Black Folks* (Institute of Black Studies, St Louis, 1975).
4. Boas, F. Introduction. *Handbook of American Indian Languages* (Government Printing Office, Washington, DC, 1911).
5. Labov, W. in *Language and Social Context* (ed. Giglioli, P.) 179–215 (Penguin, Harmondsworth, 1972).
6. Will, G. 58 boys and the larger scheme of things. *San Francisco Chronicle*, 1 January 1997.
7. Cleaver, E., *Los Angeles Times*, 31 January 1997.
8. Osterberg, T. *Bilingualism and the First School Language*. (Vasterbottens Tryckeri, Umea, 1961).
9. Taylor, H. *Standard English, Black English and Bidialectalism* (Lang, New York, 1989).
10. Piestrup, A. M. *Black Dialect Interference and Accommodation of Reading Instruction in First Grade* (Monogr. 4, Language Behavior Res. Lab., Univ. California, Berkeley, 1973).



REX FEATURES

From Cinderella's dilemma to rock slides

Jay Fineberg

A mixture of two different types of grain can undergo spontaneous stratification simply by being poured. This surprising behaviour may be significant in fields from pharmaceuticals to geology.

When Cinderella's (or Ashieppattle's, according to the brothers Grimm) wicked sisters threw her lentils into the ashes of her cooking fire, salvation came in the form of the friendly birds who helped her to sort them. But had she read the report by Makse *et al.* on page 379 of this issue¹, she might have come to a different solution to the problem of how one can go about separating two different species of intermixed granular materials. The answer may be as easy as pouring the mixture into a box. In the reported experiments, both the stratification and segregation of a mixture of two types of grain is observed to occur spontaneously as the mixture is poured into a narrow box.

Granular media show a rich variety of surprising, and at times counter-intuitive, phenomena^{2,3}. One might think, for example, that externally shaking or rotating a system composed of different types of particles would tend to induce disorder or randomize its components. In granular media the opposite occurs, as the vibration⁴ or rotation⁵ of two types of randomly mixed grain will cause them to segregate and thus increase the system's order.

A granular species can be described by the size and frictional characteristics (geometry, roughness and molecular attraction) of the grains composing it. The frictional characteristics can be characterized by the medium's 'angle of repose', the minimum slope for which gravitationally induced flow (or grain motion) can occur. Makse *et al.* prepare a random mixture of two different grain types and simply pour them into one side of a narrow container. As the slope of the 'sand-pile' formed steepens, the mixture will flow.

When the angle of repose of the grains having a larger characteristic size is greater than that of the smaller component, the flow causes spontaneous stratification of the medium to occur, and alternating layers composed of large and small particles are formed, with the smaller and 'smoother' (lower angle of repose) grains found below the larger and 'rougher' grains (Fig. 1). The authors also find that within the layers, size segregation of the grains occurs, with smaller grains tending

to be nearer the top of the pile. When the medium is chosen so that the larger grains are the 'smoother' of the two, there is no layering, and only segregation is observed.

The appearance of spontaneous grain

stratification may have important implications in both industrial and geological processes. Many industries depend on the processing and transport of granular materials. In light of this work, the assumption that, on transport, an initially well-mixed compound will remain well mixed, may need to be re-examined. The potential ramifications of stratification by means of flow may be especially important in certain applications in the pharmaceutical industry where the precise concentration of the components of a given mixture may be crucial.

The same process may also play a part in unravelling the long-standing geological puzzle of long-runout rock slides. Long-runout rock slides result from a catastrophic event such as an earthquake, explosion or meteor impact that releases a large rock mass down a mountain slope. As the name implies, the resulting landslide has the

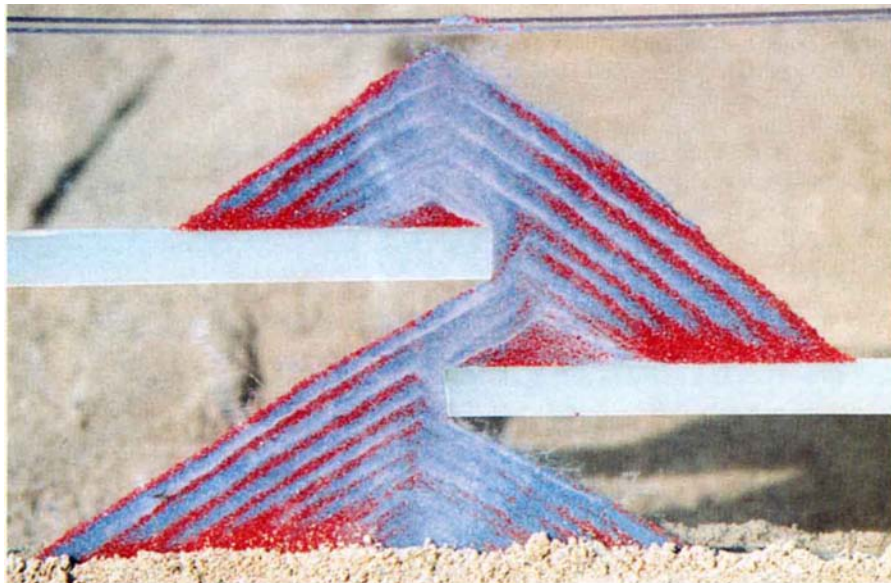


Figure 1 Spontaneous strata. Layers formed by a mixture of two types of grain, poured into a narrow container.



Figure 2 The long-runout rock slide that wiped out the town of Frank in Alberta, Canada, in 1903. After an initial one-kilometre fall, the slide travelled more than four kilometres across the valley below, and was finally stopped by the opposite slope. (Photo: National Geophysical Data Center, Boulder, CO; prepared by Allen M. Hittelman, Patricia A. Lockridge and Patrick J. Hayes.)

H. A. MAKSE

curious property that it does not stop at the bottom of the slope that drives it, but continues to slide over huge distances until arrested. Events of this nature are not uncommon, and have been known to destroy entire towns that happened to lie in their path⁶ (Fig. 2). These huge rock slides are characterized by abnormally low effective friction coefficients (defined here as the ratio of the initial fall height to the runout length), on the order of 0.1.

The origin of this low effective friction has been the subject of much speculation. Proffered explanations include an effective 'air bearing' formed by the air trapped beneath the rock mass⁷ and the 'acoustic fluidization' of a narrow zone beneath the rock layer by high intensity sound generated by the rock slide⁸. Recent simulations⁹ based on a granular model using an ensemble of smooth monodisperse disks qualitatively agree with observed flows, but many issues remain outstanding.

How, for example, can a rock slide that obviously consists of an ensemble of particles of diverse sizes and degrees of roughness be mimicked by such a simple model? An explanation may be provided by the observations of Makse *et al.* Their results suggest that the spontaneous stratification of large and small particles, with the selection of smooth smaller particles at the bottom of a given

layer, may provide an overall lubrication effect, as the best 'ball-bearings' are preferentially shifted to the bottom of the pile. In cases where acoustic fluidization occurs, this effect would tend to increase the fluidization of the medium.

As in the above example of long-runout rock slides, understanding the properties of granular materials is important. Although the subject of much active research, no underlying fundamental theoretical description of these materials yet exists. Experimental observations, such as those described by Makse *et al.*, help to provide the foundation upon which a theory for the rheology of granular materials can be based. □ Jay Fineberg is at the Racah Institute of Physics, The Hebrew University of Jerusalem, Givat Ram 91904, Jerusalem, Israel.

1. Makse, H. A., Havlin, S., King, P. R. & Stanley, H. E. *Nature* **386**, 379–382 (1997).
2. Jaeger, H. M. & Nagel, S. R. *Science* **255**, 1523–1531 (1992).
3. Jaeger, H. M., Nagel, S. R. & Behringer, R. P. *Rev. Mod. Phys.* **68**, 1259–1273 (1996).
4. Knight, J. B., Jaeger, H. M. & Nagel, S. R. *Phys. Rev. Lett.* **70**, 3728–3731 (1993).
5. Zik, O., Levine, D., Shtrikman, S. G. & Stavans, I. *Phys. Rev. Lett.* **73**, 644–647 (1994).
6. McConnel, R. G. & Brock, R. W. *Can. Dept. Inter. Annu. Rep.* 1902–1903 pt 8 (Rept. Supt. Mines, Appendix, 1904).
7. Shreve, R. L. *Science* **154**, 1639–1643 (1966).
8. Melosh, H. J. *Nature* **379**, 601–606 (1996); *Am. Sci.* **71**, 158–165 (1983).
9. Campbell, C. S., Cleary, P. W. & Hopkins, M. J. *Geophys. Res.* **100**, 8267–8310 (1995).

Developmental biology

Chick limbs, fly wings and homology at the fringe

Stephen J. Gaunt

It was from a school textbook¹ that I first understood how the limbs of different vertebrates are assumed — on the basis of their shared pentadactyl anatomy — to be derived from a common ancestral structure. All vertebrate limbs are therefore classed as being homologous. With regard to the wings of birds and insects, the textbook was clear: 'the wings of a bird and those of an insect serve the same function, but they cannot possibly be said to have originated in the same way, nor is their structure in any way comparable'. For this reason, bird and insect wings are said to be analogous, but new studies reported by Rodriguez-Esteban *et al.*² on page 360 of this issue and Laufer *et al.*³ on page 366, now lead me to think again.

The new papers report that molecular mechanisms for the formation and growth of the *Drosophila* dorsoventral wing margin are conserved — at least in part — at the margin of the vertebrate (chick) limb. In both of these structures, the dorsoventral boundary forms at the junction of the dorsal cells, which produce a protein called fringe, and the ventral cells, which do not. Fringe

protein is encoded by the *fringe* (*fng*) gene in *Drosophila*, and the *radical fringe* (*r-fng*) gene in vertebrates. In the case of the vertebrate limb, the *r-fng* expression boundary (generated as described in Fig. 1) results in the development of the apical ectodermal ridge (Fig. 2), which produces signals that are essential to maintain growth at the distal tip of the developing limb.

From work on *Drosophila*^{4,5}, it seems that fringe can serve as a boundary-determining molecule by inducing another protein, Serrate, which in turn triggers the expression (probably through the Notch receptor) of genes that are involved in wing growth and patterning on both sides of the dorsoventral boundary. Rodriguez-Esteban *et al.*² and Laufer *et al.*³ now show that the vertebrate homologues of these genes (*Serrate-2* and *Notch-1*) are expressed with *r-fng* in the apical ectodermal ridge. This indicates that the whole of this signalling pathway may be largely conserved between insects and vertebrates.

The fringe mechanism is not the only signalling pathway that is conserved in both

Drosophila wings and vertebrate limbs. But it does seem, so far, to be the pathway that is functionally the most closely conserved. Although other pathways show similarities in their relative modes of action in the *Drosophila* wing and the vertebrate limb, there are also some important differences. For example, in both structures, related members of the *hedgehog/TGF-β/patched* signalling pathways are concerned with anteroposterior patterning^{6,7}, but only in the vertebrate does this pathway lie upstream of *Hox*-gene activation⁸. Moreover, whereas expression of the *engrailed* gene confers posterior identity in the *Drosophila* wing⁹, its vertebrate homologue *Engrailed-1* is required for ventral identity in the vertebrate limb¹⁰. And *wingless* (a *Wnt* gene) may be necessary for proximodistal-axis specification in *Drosophila* appendages¹¹, but its homologue *Wnt-7a* is required for dorsal specification in the vertebrate limb^{12,13} (Fig. 1).

How are we to interpret this conservation — albeit an imperfect conservation — of the patterning mechanisms within both vertebrate limbs and insect wings? One possibility is that the two structures are homologous, both being derived from a

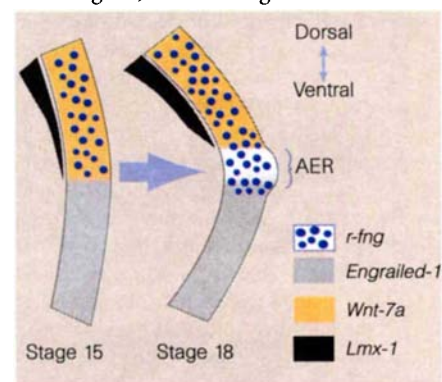


Figure 1 Dorsoventral patterning and formation of the apical ectodermal ridge (AER) during the development of vertebrate limbs involves the coordinated expression of a number of genes, shown here. During the development of *Drosophila* wings, the dorsoventral margin forms at the juxtaposition of *fng*-expressing and non-expressing cells. Rodriguez-Esteban *et al.*² and Laufer *et al.*³ now show that a vertebrate homologue of this gene (*r-fng*) functions in a similar way to regulate development of the AER at the dorsoventral margin of the vertebrate (chick) limb. Expression of *r-fng* is restricted to dorsal ectoderm partly through repression by the *Engrailed-1* expression domain in ventral ectoderm, and *r-fng* may pattern the dorsoventral boundary — as in *Drosophila* — through *Serrate* and *Notch* homologues^{2,3}. Expression of *Engrailed-1* also confines *Wnt-7a* to the dorsal ectoderm¹⁰, and *Wnt-7a* induces *Lmx-1* in the adjacent dorsal mesoderm^{12,13}. *Lmx-1* and *Engrailed-1* regulate dorsal and ventral mesodermal patterning, respectively^{10,12,13}.

common ancestral structure which itself used these pathways for developmental patterning. As noted by Laufer *et al.*³, this view is especially supported by the finding that independent signalling pathways operate together in both vertebrate limbs and insect wings. It was recently argued that insect wings are derived from ancestral gills¹⁴. So perhaps vertebrate limbs are also ultimately derived from ancestral gills, or their progenitors? Certainly, the gene-expression data obtained so far are broadly consistent with this proposal.

A significant problem in the above line of reasoning is that the various signalling pathways that have been described are not confined to insect wings and vertebrate limbs. In *Drosophila*, *fng* and *Serrate* are required for eye and leg development, and also for viability⁵. In vertebrates, *r-fng* is also expressed in branchial arches, somites, the body wall and notochord³. The *hedgehog*, *engrailed* and *Wnt* genes are also widely expressed in insect and vertebrate embryos. It is as though only a few signalling pathways were available for developmental patterning in the ancient common ancestor of insects and vertebrates, and its descendants had to use these pathways repeatedly for development of their various parts. Widespread use of the signalling pathways throughout the body means that it is not particularly surprising that several occur together in both insect wings and vertebrate limbs. So this coincidence does not, after all, provide strong evidence in favour of homology.

From the above considerations, it seems unlikely that the function of developmental genes can reliably be used as a marker of homology relationships. Otherwise, are we to conclude, for example, that the vertebrate

limb and the *Drosophila* eye (both of which use *fringe*) are homologous? Instead, during the course of evolutionary time, genetic mutations probably caused developmental genes that were already operative in one set of tissues to become ectopically activated elsewhere in the animal body. This would allow the serendipitous development of new and unrelated — that is, non-homologous — structures. These mutations could possibly lie in the control regions of the genes, and the new structures formed might then be lost or modified according to the rules of natural selection. In effect, this would mean that evolutionary change in body form was sometimes made possible as a result of homeotic-like mutations.

So, for the time being, I am inclined to the view that my old textbook is correct in its assertion that vertebrate limbs and insect

wings should not be classed as homologous — but I shall try to keep an open mind. □

Stephen J. Gaunt is in the Department of Development and Genetics, The Babraham Institute, Babraham, Cambridge CB2 4AT, UK.

1. Grove, A. J. & Newell, G. E. *Animal Biology* 6th edn (University Tutorial Press, London, 1964).
2. Rodriguez-Esteban, C. *et al. Nature* **386**, 360–366 (1997).
3. Laufer, E. *et al. Nature* **386**, 366–373 (1997).
4. Irvine, K. D. & Wieschaus, E. *Cell* **79**, 595–606 (1994).
5. Kim, J., Irvine, K. D. & Carroll, S. B. *Cell* **82**, 795–802 (1995).
6. Fietz, M. J. *et al. Development* **43s**, 51s (1994).
7. Goodrich, L. V., Johnson, R. L., Milenkovic, L., McMahon, J. A. & Scott, M. P. *Genes Dev.* **10**, 301–312 (1996).
8. Riddle, R. D., Johnson, R. L., Laufer, E. & Tabin, C. *Cell* **75**, 1401–1416 (1993).
9. Morata, G. & Lawrence, P. A. *Nature* **255**, 614–617 (1975).
10. Loomis, C. A. *et al. Nature* **382**, 360–363 (1996).
11. Campbell, G., Weaver, T. & Tomlinson, A. *Cell* **74**, 1113–1123 (1993).
12. Riddle, R. D. *et al. Cell* **83**, 631–640 (1995).
13. Vogel, A., Rodriguez, C., Warnken, W. & Izpisua Belmonte, J. C. *Nature* **378**, 716–720 (1995).
14. Averof, M. & Cohen, S. M. *Nature* **385**, 627–630 (1997).

Oceanography

Blessed dams or damned dams?

John D. Milliman

The oft-cited rationales for building dams and other river diversions (barrages, levees and channels, for example) include modulating river flow between wet and dry periods; irrigating dry and draining wet areas; producing hydro-electric power; and creating bodies of water for recreation and transport. Between 1950 and 1982, nearly 30,000 dams were constructed, more than 18,000 of which were in China¹. By the early part of this decade more than 13% of the global river flow to the sea had been dammed or diverted, and by early next century it may exceed 20% (ref. 2).

Unfortunately, even the most carefully engineered dams can create unforeseen problems, and, as with the 'Iron Gates' Dam, often in states or countries far downstream from the diversion. This dam crosses the Danube River, some 1,000 km from the river's mouths at the Black Sea. As described by Humborg and co-workers on page 385 of this issue³, it has resulted in decreased dissolved silica discharge from the Danube and, the authors surmise, in a change in the character of primary production in the Black Sea. This may be the first such cited occurrence, although decreased silica to nutrient ratios have been documented elsewhere (see, for instance, ref. 4). What makes the new paper particularly important is the possible clues it may offer to environmental problems that could occur not only in estuaries and coastal areas adjacent to dammed rivers, but also in larger bodies of water such as the Black Sea.

Better-known environmental issues highlight the diversity of deleterious effects resulting from dam construction and river diversion. For example, trapping the Nile River's waters behind the Aswan Dam clearly

mitigated the consequences of the 1980s drought in Egypt; but the dam has also had many negative downstream impacts on the river and Egypt's Mediterranean coast. Erosion seriously threatens channel and bank stability along the lower reaches of the river. Much of the sediment and water carried by the lower Nile is trapped in the extensive barrages, and in drainage and irrigation channels within the lower delta; only a little water and essentially no sediment escape to the Mediterranean. Decreased offshore catches of fish, increased shoreline erosion at the promontories, increased pollution of both land and water, and eutrophication of the once extensive coastal lagoons are some of the more obvious effects^{5,6}. Plans for additional irrigation diversions downstream from the Aswan Dam will almost certainly further reduce the quantity and quality of discharge from the Nile.

Other case histories include the Mississippi Delta, where numerous dams and extensive levee construction, coupled with increased soil conservation, have resulted in a dramatic decrease in flooding and flux of sediment to low-lying regions. Natural subsidence is therefore not compensated by sediment accumulation, resulting in rapid loss of coastal land by the expansion of coastal lagoons⁷.

In Asia, the lapse in 1988 of a water-sharing agreement between India and Bangladesh concerning the amount of Ganges River water released from the Farakka Dam (in India) resulted in a 75% reduction in the water flow to Bangladesh. In turn, this led to once fertile farmland becoming desert, as well as salt-water intrusion into the Sunderlands, a prolific mangrove forest and

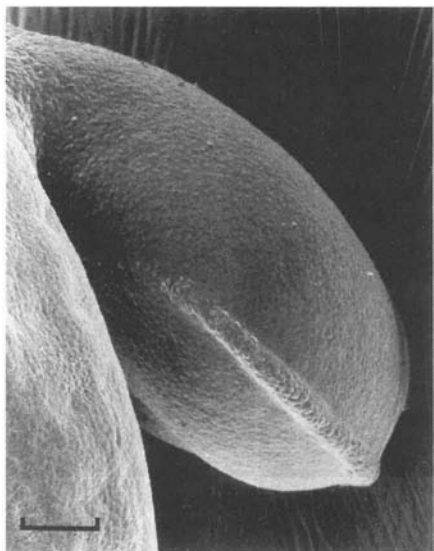


Figure 2 Scanning electron micrograph of a chick wing bud after four days of incubation. An anterior view is shown with the apical ectodermal ridge at the junction of dorsal (upper) and ventral ectoderm. Scale bar, 100 μ m. (Courtesy of J. R. Hinchliffe, University of Wales).



Figure 1 Water works on the Nile, at the Aswan Dam. Damming rivers can have environmental repercussions not only for the dammed rivers themselves and for coastal waters, but also more widely for the seas into which they flow.

natural habitat for the Bengal tiger; estimates of annual losses to Bangladesh range as high as US\$4 billion⁸. Although this dispute seems to have been resolved, the continued lack of river water downstream of the Farakka Dam will probably lead to increased groundwater extraction in the low-lying Bengal Delta, which could result in increased subsidence and thus an accelerated local rise of sea level⁹.

On the Yellow River, low rainfall and construction of upstream dams (combined with increased soil conservation) have led in recent years to a nearly 50% drop in both water and sediment discharge, and in 1995 the river was dry for four months. Continued low sediment discharge puts at risk the river's delta, the site of major petroleum fields.

To date, most of the documented effects of dammed rivers have involved decreased water or sediment flux to downstream portions of river basins. More subtle, but perhaps with long-term implications, will be changes in the chemical and biological environments of coastal waters such those as described by Humborg *et al.*³. Decreased sediment loads due to increased soil conservation and river diversion, together with increased nutrient loading from agricultural fertilizers, for example, might be a major cause of increased eutrophication and severely reduced levels of oxygen in North American and European coastal waters^{4,10,11}.

Humborg *et al.* show that decreased flux of dissolved silica from the Danube, which

contributes about 70% of the freshwater input into the Black Sea, has resulted in decreased silica concentrations in the surface layer throughout the sea. They further argue that a consequence has been a shift in the composition of phytoplankton species from siliceous to non-siliceous. But decreased discharge of river water also may have a considerable influence on the circulation within the Black Sea. If all the other major rivers leading

to the Black Sea — the Dnieper, Don, Rioni and Sakarya — were completely dammed, for example, the decreased freshwater flux to the Black Sea could diminish the mixed freshwater layer^{12,13}, perhaps leading to a shoaling of deeper waters rich in hydrogen sulphide and possible environmental repercussions for the sea and its bordering countries. In turn, decreased flow from the Black Sea and the dammed Nile may affect the freshwater budget in the eastern Mediterranean¹⁴.

As scientists develop a better understanding of land-sea interactions, we should be able to predict more accurately how new river diversions can be built to minimize undesirable side effects. For some dams the question may not be if a dam is built, but how it is built. □

John D. Milliman is in the School of Marine Science, Virginia Institute of Marine Science of the College of William and Mary, Gloucester Point, Virginia 23062, USA.

1. van der Leeden, F., Troise, F. L. & Todd, D. K. *The Water Encyclopedia* (Lewis, Chelsea, MI, 1990).
2. Pierce, F. *New Scientist* 4 May, 36–39 (1991).
3. Humborg, C., Ittekkot, V., Cociasu, A. & v. Bodungen, B. *Nature* **386**, 385–388 (1997).
4. Turner, R. E. & Rabalais, N. *Nature* **368**, 619–621 (1991).
5. Kempe, S. *Mitt. Geol.-Paläontol. Inst. Univ. Hamburg* **74**, 81–114 (1993).
6. Stanley, D. J. *Mar. Geol.* **129**, 189–195 (1996).
7. Wells, J. T. & Coleman, J. M. *Estuar. Coast. Shelf Sci.* **25**, 111–125 (1987).
8. Cooper, K. J. *Washington Post* 21 Dec., A-20 (1996).
9. Milliman, J. D., Broadus, J. M. & Gable, F. *Ambio* **18**, 340–345 (1989).
10. Kempe, S., Pettine, M. & Cauwet, G. in *Biogeochemistry of Major World Rivers* (eds Degens, E. T., Kempe, S. & Richey, J. E.) 169–211 (Wiley, Chichester, 1991).
11. Degens, E. T., Kempe, S. & Richey, J. E. (eds) in *Biogeochemistry of Major World Rivers* 323–347 (Wiley, Chichester, 1991).
12. Murray, J. W., Top, Z. & Ozsoy, E. *Deep-Sea Res.* **38**, S663–S689 (1991).
13. Tugrul, S. *et al. Nature* **359**, 137–139 (1992).
14. Rohling, E. J. & Bryden, H. L. *J. Geophys. Res.* **97**(C7), 11191–11198 (1992).

Physiological ecology

Leaning trees on sloping ground

R. McNeill Alexander

In forests on hillsides, the trunks of the understorey trees may lean outwards from the slope, although the canopy trees are more vertical (Fig. 1a, overleaf). Two main theories have been put forward to explain this phenomenon — that it is caused either by the wind or by disturbances such as landslides. But in *Proceedings of the Royal Society*, Ishii and Higashi¹ offer a new explanation, and they present data that are consistent with it from an evergreen forest on an island off the southern tip of Japan.

The principle is very simple. Plants depend on the light energy that they capture by photosynthesis. The light comes from the Sun, and it gets progressively dimmer as it descends through a forest, from the canopy to the ground. The lines of equal light inten-

sity (Fig. 1a) run parallel to the ground², so to obtain as much light as possible at its crown, a tree of given trunk height should grow at right angles to the ground — on level ground it should grow vertically, but on a slope it should lean outwards. However, the greater the angle of lean, the stronger the trunk of the tree needs to be, and the more firmly must its roots be fixed in the ground if it is not to snap or fall over. With limited resources of energy and materials, a leaning tree cannot grow as tall a trunk as if it were vertical, so the optimal angle for a tree on a slope is neither vertical nor perpendicular to the ground, but somewhere between these extremes.

Optimality arguments are common in biology, and many of them depend on the

very powerful optimizing process of evolution by natural selection³. But whereas most of these models calculate quantities (such as energy gain) that are presumed to be correlated to survival or reproductive success, Ishii and Higashi¹ set out to calculate survival itself. To do this, they have formulated a remarkably elaborate model. It goes like this.

Given the length of an understorey tree's trunk and the angle to the vertical at which it is tilted, simple trigonometry gives the height of the crown above the ground. An assumption about the amount of light that is blocked out by taller trees allows the intensity of the light falling on the leaves to be calculated. The quantity of leaves can be estimated from the size of the trunk, and a well-known relationship makes it possible to calculate the rate of photosynthesis from the light intensity.

Assume that if there were no need to allow for its weight-supporting function, the radius of the trunk would be one per cent of its length. Now consider the problem of weight support — taking account of the angle of the tree to the vertical, as well as its height — and calculate how much additional wood is needed in the lower parts of the trunk. From the total volume of wood estimate the metabolic rate of the tree, and subtract this from the photosynthetic rate to obtain the rate at which energy and materials become available for growth.

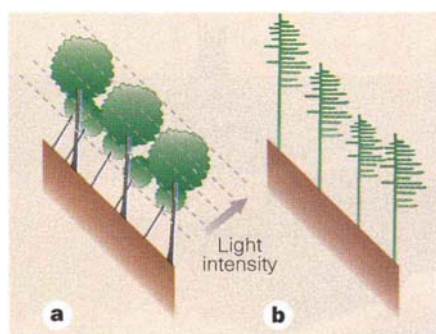


Figure 1 a, Ishii and Higashi¹ have studied why understorey trees that grow on slopes often incline their trunks downwards. They conclude that there is an optimal angle of growth for maximal survival, and that trees at an angle can get brighter illumination for the leaves by shortening the distance from the canopy surface. Lines of equal light intensity are shown, running parallel to the ground. b, The same effect is achieved by trees that have more branches on the downhill side.

Assume that the lower this rate is, as a fraction of the rate of photosynthesis, the more likely the tree is to die. Calculate the probability of survival from seedling to mature tree by integrating over time, taking account of the rate of growth and the mortality at every stage. Finally, for each angle of slope, calculate this probability of survival

for a range of trunk inclinations.

Ishii and Higashi used this procedure to show that, as expected, for any given angle of slope there is an angle of inclination of the trunk that maximizes survival. With environmental and physiological parameters that they select as being plausible for shade-sensitive understorey trees, the authors predict that the optimal angle to the vertical is 0° on level ground, about 10° on a 20° slope, and about 40° on a 50° slope. For shade-tolerant trees, the effect is less marked as light intensity has less effect on their photosynthesis. In the forest that they studied, Ishii and Higashi found the predicted relationship between slope and trunk inclination for one understorey species, *Rhododendron tashiroi*, but there was no significant relationship for another species that was more tolerant of shade.

Surprisingly, it is not clear whether the authors suppose that trees tend to grow towards the light. They find that *Rhododendron* trunk inclination on steep slopes has high variance, and argue that this is consistent with their prediction that, on steep slopes, a range of angles around the optimum will give almost equal survival. This suggests that they envisage seedlings growing indiscriminately at a number of different angles, with only the appropriately angled trees surviving to maturity. ►

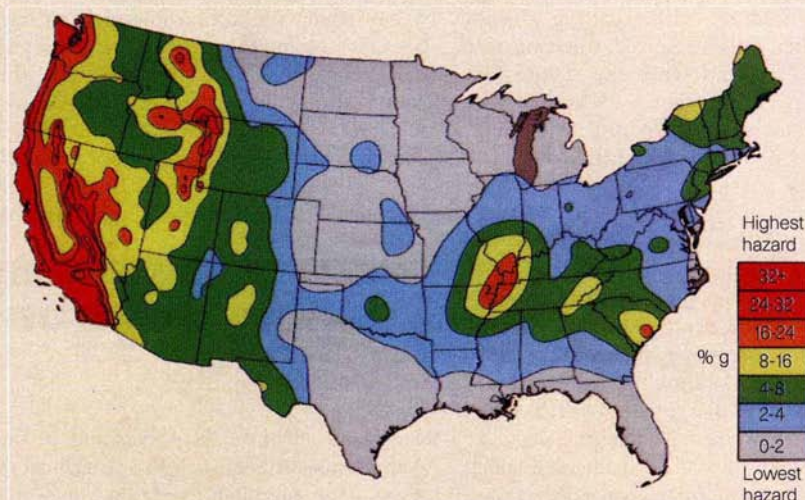
Earthquakes

Shaking in the USA

A new set of maps, released by the US Geological Survey, show the latest estimates of earthquake shaking hazards in the United States. Along with the well-known hazards in California, they reveal an increased awareness of the threat of earthquakes in the Pacific Northwest (associated with the Cascadia subduction zone), the central United States (stemming, for example, from the New Madrid fault zone), and regions affected by the Yellowstone hot spot (Wyoming, Idaho and Montana).

The colours on the simplified map shown here represent the level of horizontal shaking that has a 1-in-10 chance of being exceeded in a 50-year period. The potential shaking is represented as a percentage of *g* — gravitational acceleration at the Earth's surface. (So a vertical acceleration of 100 would be sufficient to throw objects into the air.)

The national maps have been produced since 1948, and are used in the design of buildings, highways, utilities and bridges to help to produce structures that will withstand the amount of shaking predicted to be likely for a particular region. The latest versions have been generated by using



geological slip rates from an unprecedented number of faults (nearly 500), and can be viewed at <http://geohazards.cr.usgs.gov/eq/>.

A related development is the publication of a special issue of *Seismological Research Letters* (68, January/February 1997), which presents the methods and data behind the quantification of ground shaking from earthquakes — a major aspect of estimating seismic hazard. The 12 papers describe the science behind seismic hazard assessment in a form that will be of use to public officials as well as seismic engineers.

The estimation of expected ground motion at a particular site depends not only on the magnitude of the potential earthquake and the distance of the site from the epicentre, but also on the type of faulting (orientation of the rupture plane) and conditions at the site concerned (for instance whether it consists of hard rock or unconsolidated sediment). The estimates are also calculated for a range of frequencies, because the sensitivity of a building or other structure to shaking at a given frequency will depend on its size and style of construction.

John VanDecar

Ishii and Higashi ignore the possibility that a vertical trunk may bear an asymmetrical crown, but such crowns have been observed on hillsides¹. This may result from the more rapid growth of well-lit than shaded branches on the same tree², so it could be explained without any optimality model. However, a modification of Ishii and Higashi's model that considers the growth and mortality, not of the complete tree but of individual branches (as in ref. 5), could explain a casual observation — larch (*Larix*) trees on steep slopes near my home seem to

have more branches on the downhill side of the trunk (Fig. 1b). □

R. McNeill Alexander is in the Department of Biology, University of Leeds, Leeds LS2 9JT, UK.

1. Ishii, R. & Higashi, M. *Proc. R. Soc. Lond. B* **264**, 133–140 (1997).
2. Horn, H. S. *The Adaptive Geometry of Trees* (Princeton Univ. Press, NJ, 1971).
3. Alexander, R. M. *Optima for Animals* 2nd edn (Princeton Univ. Press, NJ, 1996).
4. Halle, E., Oldemann, R. A. A. & Tomlinson, P. B. *Tropical Trees and Forests: an Architectural Analysis* (Springer, Berlin, 1978).
5. Sorrenson-Coothern, K. A., Ford, E. D. & Sprugel, D. G. *Ecol. Monogr.* **63**, 277–304 (1993).

DNA topoisomerases

New break for archaeal enzyme

James C. Wang

DNA topoisomerases have been dubbed the 'magicians' among the enzymes that catalyse cellular transformations — in their presence, individual DNA strands and double helices can pass through each other as if there were no physical barriers in-between¹. Topoisomerases probably entered the stage of evolution when the increasing length of DNA, or the formation of duplex DNA rings, created topological problems that needed to be resolved for propagation of the genetic material: for example, unravelling of the complementary DNA strands during replication; disentanglement of duplex DNA molecules during chromosomal condensation and segregation; and prevention of excessive supercoiling of the DNA template during transcription. Several mechanistically distinct DNA topoisomerases are known, and on page 414 of this issue, Bergerat *et al.*² describe a fascinating new subfamily, through their study of an unusual topoisomerase from the archaeon *Sulfolobus shibatae*.

Type I DNA topoisomerases transiently break one DNA strand at a time for the enzyme-mediated passage of another strand, whereas the type II enzymes break both strands of a DNA double helix for the passage of a second double helix. Both types use the same basic chemistry in their transient breakage of DNA — an enzyme tyrosyl group attacks and breaks a DNA backbone bond, then becomes linked to one side of the break by forming a phosphotyrosine bond, leaving a deoxyribose hydroxyl group on the other side of the break (Fig. 1). Reversal of this reaction disrupts the enzyme–DNA covalent link and rejoins the broken DNA strand.

The type I DNA topoisomerases fall into two subfamilies with distinct amino-acid sequences and biochemical characteristics. By contrast, all of the type II DNA topoisomerases (ranging from those encoded by viruses and bacteria to those encoded by plants and mammals), were thought to form a single subfamily. Initially, there was no indi-

cation that the type II DNA topoisomerase that was purified by Bergerat *et al.*³ from *S. shibatae* differed greatly from the other type II enzymes: it is ATP-dependent and it shows the topological signature of all such enzymes in that it removes supercoils two at a time. But partial peptide-sequencing and gene cloning showed that there is little sequence homology between the two subunits of the *S. shibatae* enzyme and polypeptides of the known type II topoisomerases. So Bergerat *et al.*² named this enzyme *S. shibatae* DNA topoisomerase VI, following the tradition of naming the type II enzymes by even numbers.

The identification of a new type II DNA topoisomerase that has a distinct primary structure offers one plausible answer to a recent puzzle. In bacteria and eukaryotes, at least one type II DNA topoisomerase is required for viability¹. Sequencing of the entire *Methanococcus jannaschii* genome, however, initially suggested that this archaeon might not possess such an enzyme⁴. Because bacterial and eukaryotic type II DNA topoisomerases are indispensable in the final segregation of intertwined pairs of fully duplicated chromosomes, the apparent absence of such an enzyme in the archaeon could be attributed to a parallel pathway that uses a type I DNA topoisomerase to unlink incompletely duplicated chromosomal pairs, yielding DNA molec-

ules with small, single-stranded gaps in them. But the *S. shibatae* DNA topoisomerase VI offers an alternative interpretation — that a hitherto unidentified type II topoisomerase might be involved in chromosomal segregation in *M. jannaschii*. Indeed, genes with nucleotide sequences that are homologous to the structural genes of *S. shibatae* DNA topoisomerase VI have now been identified in *M. jannaschii*².

Particularly significant is the finding that the new archaeal type II topoisomerase has eukaryotic homologues that are involved in meiotic events in yeast. The initiation of meiotic recombination in the budding yeast *Saccharomyces cerevisiae* involves a protein that makes double-stranded cuts, and becomes covalently attached to the 5' ends of the broken DNA^{5–7}. This 5'-linked polypeptide has been identified as Spo11 (ref. 8), and Bergerat *et al.*² have found that it is a homologue of the A subunit of DNA topoisomerase VI. A yeast protein that is homologous to the B subunit of *S. shibatae* could, therefore, be acting in conjunction with Spo11 to introduce meiosis-specific DNA breaks, and Bergerat *et al.*² suggest that Hsp82 might be such a subunit.

One is then confronted with the problem that, whereas a topoisomerase breaks and rejoins DNA strands, the formation of meiosis-specific DNA breaks is normally followed by resection of the new 5' ends to yield DNA molecules with long, single-stranded 3' tails. Studies of drugs and polypeptide toxins, many of which stabilize the topoisomerase–DNA covalent intermediates, indicate a possible solution. Presumably, a topoisomerase-generated covalent complex (stabilized by protein modification or interaction with another protein) can be processed further to give DNA molecules with single-stranded ends. Equally plausible is that the Spo11-containing meiotic activity might have diverged sufficiently from the type II topoisomerases so that its main function is to cleave DNA duplexes but not to rejoin DNA ends.

The DNA topoisomerases and many of the site-specific recombinases have been known to be mechanistically similar for some time. Now it seems that a topoisomerase-type activity might also be instru-

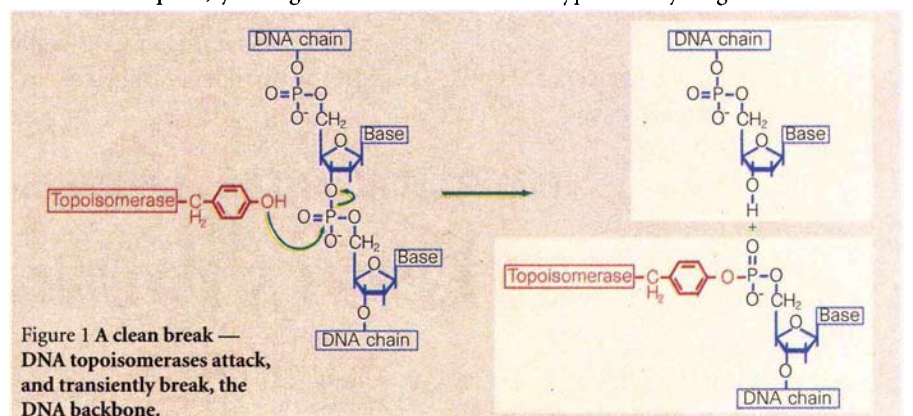


Figure 1 A clean break — DNA topoisomerases attack, and transiently break, the DNA backbone.

mental in meiotic recombination. Nature has once again demonstrated its unfailing ability to create enlightening variations of a beautiful theme. □

James C. Wang is in the Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

1. Wang, J. C. *Annu. Rev. Biochem.* 65, 635–692 (1996).

2. Bergerat, A. *et al. Nature* 386, 414–417 (1997).
3. Bergerat, A., Gadelles, D. & Forterre, P. *J. Biol. Chem.* 269, 27663–27669 (1994).
4. Bult, C. J. *et al. Science* 273, 1058–1072 (1995).
5. de Massy, B., Rocco, V. & Nicolas, A. *EMBO J.* 14, 4589–4598 (1995).
6. Liu, J., Wu, T.-C. & Lichten, M. *EMBO J.* 14, 4599–4608 (1995).
7. Keeney, S. & Kleckner, N. *Proc. Natl Acad. Sci. USA* 92, 11274–11278 (1995).
8. Keeney, S., Giroux, C. N. & Kleckner, N. *Cell* 88, 375–384 (1997).

Active galaxies

Hot winds stay in the plane

Fred Hamann

Astronomers first detected high-speed plasma outflows from quasars in the late 1960s, just a few years after the quasars themselves were recognized as extremely distant and luminous objects. Since then we have grappled with basic questions about these winds. What drives them? What is their geometry? What is the source of outflowing gas? How much gas is involved and what is its ionization state and chemical composition? New progress on these questions, stimulated by observations using satellites and large ground-based telescopes, was discussed at a meeting in Pasadena in February*.

Current theory holds that quasars are massive black holes (up to about a billion times the mass of our Sun) fuelled by gaseous accretion disks. The material inside the disks becomes extremely hot and shines brightly as it spirals down the gravitational funnel towards the black hole. The result is an object less than 0.01 light years across that typically outshines the entire galaxy around it. Out of these exotic environments come outflows of hot, ionized gas, driven up to speeds that sometimes exceed 50,000 km s⁻¹ (nearly 0.2 times the speed of light). We measure the amount and speed of the outflowing plasma via ultraviolet absorption lines that are wavelength-shifted by the expansion.

A fundamental characteristic of quasar outflows is the amount of each ion present. For years we have naively assumed that shallow absorption lines result from relatively small numbers of absorbing ions, but one of the points stressed in Pasadena is that what you see is not necessarily what you get. Some of the quasar light can, for example, leak unabsorbed through holes in the wind, and thereby fill in the bottoms of what would otherwise be deep absorption troughs. Our resulting tendency to underestimate the number of absorbing ions can be corrected in some cases by comparing different absorption lines of the same ion — where the true relative line strengths are known from atomic physics. High-resolution spectra obtained with large, ground-based telescopes in the past few years have allowed us to apply this comparative analysis to several

quasars and obtain the first reliable measures of the wind material^{1–3}.

The best estimates of the total outflowing mass remain close to old accepted values — roughly 0.1 solar masses, varying by at least a factor of ten between quasars. But new satellite X-ray observations^{4,5} suggest that the ions we measure in the ultraviolet, such as C³⁺, are just trace constituents in a highly ionized plasma dominated by ions like O⁺⁶ and O⁺⁷. The X-ray data do not provide accurate velocity information, so it is not yet clear whether the enormous amounts of highly ionized gas are part of the outflows revealed by the ultraviolet lines. If they are, the total flow masses would be at least 100 to 1,000 times larger than previous estimates.

Perhaps the most interesting results discussed in Pasadena come from measurements of the polarization of quasar light^{6–8}. The light emitted by quasars becomes polarized if it scatters off surrounding material, possibly dust or free electrons far above the accretion disk. The measured spectrum can be divided into the polarized (scattered) and unpolarized (direct) components to provide several 'views' of the same quasar (Fig. 1). We therefore have an indirect probe of the geometry of the outflow.

The growing opinion among several research groups is that the flows are concentrated near the surface of the accretion disk. This result fits nicely into schemes that attempt to explain differences in the measured properties of quasars by their orientation. In particular, quasar accretion disks are believed to be just the inner extensions of much larger and thicker disks or tori that have galactic dimensions (thousands of light years)

and are opaque to most radiation. If we view one of these disk systems edge-on, the central quasar is obscured, but still contributes polarized flux by scattering light off the tenuous debris above and below the disk plane. The observed flux in this case is highly polarized. The same system viewed pole-on would reveal the unobscured quasar shining brightly, with relatively little contribution from scattered, polarized light. Several studies have now confirmed that quasar spectra with the signatures of an outflow also tend to be highly polarized. Evidently, when we view a quasar/disk system nearly edge-on (high polarization), we are also looking through the outflow, so quasar winds lie close to the disk plane.

Most theoretical models now incorporate this geometry. There is also a growing consensus that light from the bright central quasar both ionizes and accelerates the gas. The models remain significantly divided, however, about the source of outflowing material. Some invoke the accretion disk as a gas reservoir, whereas others use material ablated off stars near the quasar. Resolving this issue is critical, but it is not yet clear whether any existing model can explain the wealth of data.

In his meeting summary, Colin Norman reminded us of our ultimate goals. We study quasar outflows to learn more about these peculiar objects and their exotic environments. But several studies are now also connecting this work to broader problems in astronomy. For example, we know that all of the heavy elements from carbon up are synthesized from hydrogen and helium in stars. Estimates of the elemental abundances near distant quasars can therefore constrain the amount of star formation and chemical processing in the early Universe. New results from quasar winds indicate that the heavy elements are often more abundant in these environments than they are in the Sun^{9–11}. Apparently, the star formation is rapid — most of the gas initially in the centres of massive galaxies (where quasars reside) collapses into stars just a few billion years after the Big Bang. The gas becomes enriched by frequent supernova explosions, which are the spectacular fate of short-lived massive stars. This evolution is frenzied compared with what occurs in outer galactic regions, such as the solar neighbourhood of our own Milky Way. □

Fred Hamann is in the Center for Astrophysics and Space Science, University of California at San Diego, La Jolla, California 92093-0424, USA.

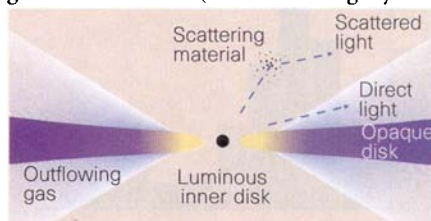


Figure 1 Quasar geometry. The dashed lines are light rays seen by an observer to the right. The scattered light is highly polarized, and the direct light is attenuated by the outflow, which also gives it blue-shifted absorption lines.

* Mass Ejection From Active Galactic Nuclei, Carnegie Observatories, Pasadena, California, USA, 19–21 February 1997.

1. Petitjean, P., Rauch, M., & Carswell, R. F. *Astron. Astrophys.* 291, 29–58 (1994).
2. Barlow, T. A. & Sargent, W. L. W. *Astron. J.* 113, 136–143 (1997).
3. Hamann, F., Barlow, T. A., Junkkarinen, V. T. & Burbidge, E. M. *Astrophys. J.* (in the press).
4. Mathur, S., Elvis, M. & Singh, K. P. *Astrophys. J.* 455, L9–L12 (1995).
5. Green, P. J. & Mathur, S. *Astrophys. J.* 462, 637–642 (1996).
6. Hines, D. C. & Wills, B. J. *Astrophys. J.* 448, L69–L72 (1995).
7. Goodrich, R. W. & Miller, J. S. *Astrophys. J.* 448, L73–L76 (1995).
8. Cohen, M. *et al. Astrophys. J.* 448, L77–L80 (1995).
9. Turnshek, D. *et al. Astrophys. J.* 463, 110–119 (1996).
10. Korista, K., Hamann, F., Ferguson, J. & Ferland, G. *Astrophys. J.* 461, 641–656 (1996).
11. Hamann, F. *Astrophys. J. Suppl. Ser.* (in the press).

NATURE

100 YEARS AGO

The majority of deep-sea species are mud eaters; some are of gigantic size; some are armed with peculiar tactile, prehensile, and alluring organs; some are totally blind, whilst others have large eyes and are provided with a kind of dark lantern for the emission of phosphorescent light. The deep-sea fauna does not represent the remnants of very ancient faunas, but has rather been the result of migrations from the region of the mud-line in relatively recent geological times. The *Challenger* investigations show that species are most abundant in the shallow waters near land, decreasing in numbers with increasing depth, and especially with increasing distance from continental land. This is true as a general rule, especially of tropical waters, but in polar regions there are indications of a more abundant fauna in depths of 50 to 150 fathoms than in shallower water under 50 fathoms. The various points touched upon regarding the distribution of marine organisms, might be explained on the hypothesis that in early geological times there was a nearly uniform high temperature over the whole surface of the globe, and a nearly uniformly distributed fauna and flora; and that with the gradual cooling at the poles, species with pelagic larvæ were killed out or forced to migrate towards the tropics, while the great majority of the species which were able to survive in the polar areas were those inhabiting the mud-line. The uniform physical conditions here referred to might be explained by adopting the views of Blandet as to the greater size and nebulous character of the sun in the earlier ages of the earth's history.

From *Nature* 25 March 1897.

50 YEARS AGO

The advantages of hydrogen cooling for large high-speed electrical machines have not up to now been exploited in Great Britain. Several manufacturers are at present engaged on the construction of turbo-alternator units employing hydrogen as a cooling agent. In a report on work and progress in 1946 issued by the Metropolitan-Vickers Electrical Co., Ltd., it is stated that a 60,000 kW., 3,000 r.p.m. set is approaching completion and will be ready for testing shortly. — "Hydrogen-cooled turbo-alternators."

From *Nature* 29 March 1947.

Evolutionary ecology

The right size for a mammal

Andy Purvis and Paul H. Harvey

Elephants live slow lives, voles fast lives — vole populations may pass through 50 generations while an elephant grows up. Many facets of mammalian life history combine in this slow-fast continuum: species with long generation times produce small litters of large offspring who wean late, mature late and live longer after maturity¹. Why do these components of life history co-vary as they do? In a paper in *American Naturalist*², Kozłowski and Weiner propose an elegant and comprehensive model of life-history evolution for animals, such as mammals, that stop growing when they reach reproductive maturity. Their model accords with the comparative patterns and may, as a bonus, resolve another knotty problem — the skewed frequency distribution of body sizes among species, whereby far more species are small than large (Fig. 1).

The study of mammalian life-history evolution gained impetus from the demonstration of impressively tight scaling relationships between life-history attributes and body size among species³ — large species lead slow lives and small species fast ones. However, the fast-slow continuum cannot be just a side-effect of natural selection on size because it exists independently of size differences: voles are fast, for example, but bats are slow. Empiricists and theoreticians have batted alternative explanations back and forth, setting each other ever more demanding challenges^{4,5}. Ironically, in Kozłowski and Weiner's explanation, the interspecies scaling relationships that sparked the whole debate turn out to be red herrings.

Kozłowski and Weiner's new model assumes that, within a species, the rates of energy assimilation, respiration and mortality scale with body size according to simple power laws, yielding straight lines on double-logarithmic plots. Further, differences in ecology cause both the slopes and the intercepts of these lines to vary among related species (Fig. 2). Within a species, these six

parameters — the three slopes and three intercepts — are facts of life. Evolution responds to them by optimizing the timing of the switch from investing energy in growth to investing in reproduction⁶. Optimal size varies among species because of ecological differences. But various combinations of parameter values can yield the same optimum adult size for different species and, when this happens, one species matures later than the other and lives longer afterwards: the model captures the slow-fast continuum.

For each species, the optimal size corresponds to a particular growth rate, age at maturity, adult mortality rate and so on. In simulations, these life-history traits scale as power functions of adult body size among species. The exact scalings depend upon the distributions of the six ecologically determined parameters that went into the model but, for some realistic choices at least, the likeness to observed interspecies scalings is uncanny. However, the scaling relationships carry no functional significance — the slopes are not the same as within-species slopes, but are statistical consequences of body-size optimization. Another finding deserves special mention. In Kozłowski and Weiner's simulations, each ecological parameter is assumed to be normally distributed among species. The resulting distribution of optimal body sizes, however, is skewed to the right even when body size is logarithmically transformed. Actual body size distributions are often this shape too⁷, and a long search for the reason has hitherto proved inconclusive.

Kozłowski and Weiner are not the first to tackle life-history evolution by combining optimality theory with body-weight scaling of life-history variables. Charnov^{4,8} produced a model having many of the same predictions which tightly couples size to age by assuming a growth law, with all species growing along parallel trajectories. His model, which also features optimization of adult size, successfully recovers both the slow-fast continuum and the scaling relationships between life history and adult size, and has led to insights into bird life-history too⁹. However, tests of the model against real data have found flaws: most importantly, body weight at weaning is not, as Charnov assumes, a constant proportion of adult weight⁵. As Kozłowski and Weiner point out, the assumption is illogical: it precludes the existence of an optimum size, because delaying reproduction would incur the double whammy of lower survivorship to maturity and lower fecundity once mature.

Kozłowski and Weiner's model is an advance over Charnov's, but how good is it?

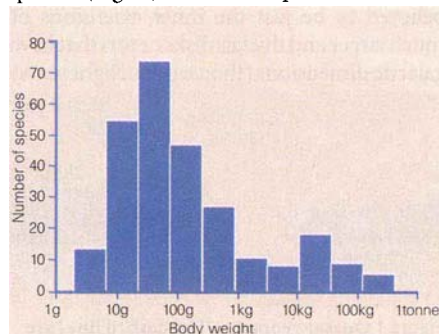


Figure 1 Distribution of adult body weights among Palearctic terrestrial mammals, which is skewed so that it has a long right-hand tail¹³. The shape is typical of other plots for other areas.

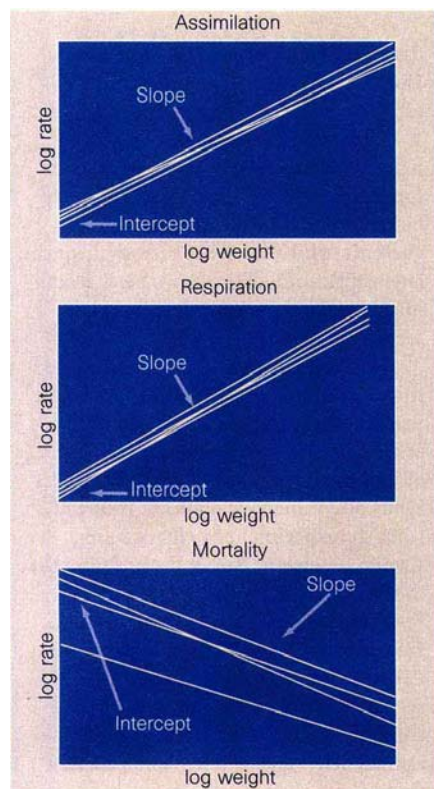


Figure 2 The basis of Kozlowski and Weiner's model² of life-history evolution for animals that stop growing when they reach reproductive maturity. Within a species, the rates of energy assimilation, respiration and mortality scale with body weight. On the graphs, each line represents a different species. Across species, the slopes and intercepts of the scaling relationships vary because of ecological differences, and are normally distributed within each part of the figure. The consequence is that different species may have different optimal body sizes.

Although the fast-slow continuum is robust to the precise choice of values assigned to ecological parameters, the means and especially the variances critically affect the interspecies scaling. Unfortunately, because we do not know what values are appropriate, quantitative tests are not yet feasible. The model's view of how mammals grow nonetheless permits one qualitative test. Whereas Charnov assumed that the scaling of growth rate on body size would be the same within species as among them, Kozlowski and Weiner expect within-species scalings to be less steep. We analysed published growth equations¹⁰ between weaning and maturity in 31 species for which life-history timings⁵ are available (Fig. 3). In every case, the within-species scaling is less positive than the among-species relationship, just as Kozlowski and Weiner (but not Charnov) predict.

The new model has important implications. The values of cross-taxonomic scaling coefficients and exponents that relate history variation to body size have attracted considerable interest (see ref. 11). But, if Kozlowski and Weiner are right, such values are all but

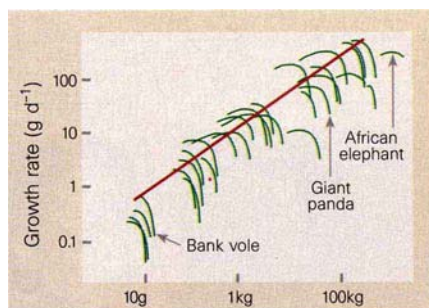


Figure 3 A test of Kozlowski and Weiner's model². Red line: interspecies scaling of maximum growth rate versus adult weight (slope estimated from independent comparisons¹¹). Green lines: for each species, how growth rate scales with mass after weaning (the right-hand continuation of lines to the body-size axis, when growth rate slows as maturity is approached, is not shown). Note that the interspecies line is more positive than any part of any within-species line.

meaningless for life history at least. Little wonder, then, that the generation of functional hypotheses from scaling relationships has not proved fruitful. Second, the peak in the body-size distribution is not an optimum for all mammals; rather, each species' optimum depends upon its combination of ecological parameters, and not all combinations are equally likely. Finally, the model casts light on why it has proved so hard to correlate life-history differences with ecology: related species with differing ecologies are expected to be different sizes, and life-history variation correlated with size differences is commonly factored out in comparative studies. By viewing interspecies scalings of life-history variables on body weight as epiphenomena and ecological differences as ultimate causes, Kozlowski and Weiner's model focuses attention on another key issue currently enjoying a renaissance — the evolution of body size itself¹².

Andy Purvis is in the Department of Biology, Imperial College, Silwood Park, Ascot SL5 7PY, UK. Paul H. Harvey is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

1. Harvey, P. H., Read, A. F. & Promislow, D. E. L. *Oxf. Surv. Evol. Biol.* **6**, 13–31 (1989).
2. Kozlowski, J. & Weiner, J. *Am. Nat.* **149**, 352–380 (1997).
3. Western, D. *Afr. J. Ecol.* **17**, 185–204 (1979).
4. Charnov, E. L. *Life History Invariants: Some Explorations of Symmetry in Evolutionary Ecology* (Oxford Univ. Press, 1993).
5. Purvis, A. & Harvey, P. H. *J. Zool.* **237**, 259–283 (1995).
6. Perrin, N. & Sibley, R. M. *Annu. Rev. Ecol. Syst.* **24**, 379–410 (1993).
7. May, R. M. in *Diversity of Insect Faunas* (eds Mound, L. A. & Waloff, N.) 188–204 (Blackwell, Oxford, 1978).
8. Charnov, E. L. *Proc. Natl Acad. Sci. USA* **88**, 1134–1137 (1991).
9. Owens, I. P. F. & Bennett, P. M. *Proc. R. Soc. Lond. B* **261**, 227–232 (1995).
10. Zullinger, E. M., Ricklefs, R. E., Redford, K. H. & Mace, G. M. *J. Zool.* **65**, 607–636 (1984).
11. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, 1991).
12. Jablonski, D. *Nature* **385**, 250–252 (1997).
13. Purvis, A. & Harvey, P. H. *Symp. Zool. Soc. Lond.* **69**, 159–174 (1996).

Daedalus

It's all in the mind

Compulsory games in British schools are declining. A public discussion of the matter a few years ago showed a curious distribution. Some adults recalled football and other sports as the high point of their school lives; others remembered them with loathing and rage. In this connection, Daedalus notes the claim that vigorous exercise releases endorphins into the body, stimulating the pleasure receptors. Clearly, he says, this only happens to some people. They are the extroverts who enjoy physical exertion. The introverts who detest it gain no pleasurable sensation from exercise at all; for them it is just draining.

So, says Daedalus, British education could be greatly improved by giving each pupil a simple exercise test, while taking a blood sample for endorphin analysis. The scholars could thus be neatly divided from the sporting types, and treated accordingly. But why does this division exist at all? Daedalus sees evolution at work. Our primitive ancestors often had physical exertion forced upon them: as in tribal warfare, or hunting or running away from mammoths and other dangerous creatures. Evolution wisely encouraged them with the endorphin mechanism. But as human intelligence advanced, some hominids began to succeed by cunning rather than violence. They lost the endorphin reward for muscle metabolism. Indeed, muses Daedalus, they may even have begun to be rewarded by brain metabolism instead.

So DREADCO psychologists are looking for the hitherto unremarked phenomenon of 'thinker's high'. They are seeking out compulsive intellectuals of the type of Stanislaw Ulam, whose daughter complained "All my father does is think, think, think! Nothing but think!". They are challenging them with fearsomely difficult problems and paradoxes, sampling their blood for analysis. The ultimate result could be a new form of careers advice, identifying the nation's natural thinkers for suitable education and employment.

Most people, however, are less extreme. They gain some reward from both mental and physical exertion. The balance between them may even be adjustable. Ritaline, the amphetamine used to control hyperactivity in children, may work by boosting mental rewards while damping physical ones. Food dyes and additives, such as tartrazine, have been blamed for causing hyperactivity in the first place. They may bias endorphin release the other way. They could be useful for snapping absent-minded professors back into the real world.

David Jones

Chien-Shiung Wu (1912–97)

Experimental physicist,
co-discoverer
of parity violation

Chien-Shiung Wu died of a stroke on 16 February 1997. Forty years earlier, on 15 February 1957, *The Physical Review* (105, 1413) published the historic paper, "Experimental test of parity conservation in beta decay", by C. S. Wu of Columbia University and E. Ambler, R. W. Hayward, D. D. Hoppes and R. P. Hudson of the National Bureau of Standards (NBS). In that paper, through investigations on the beta-decay of cobalt-60, they established for the first time the non-conservation of parity (P) and the violation of particle-antiparticle conjugation symmetry (C) in physics, both of which had been taken for granted as basic laws of nature.

Parity in physics means right-left symmetry. That is, take two systems which are mirror images of each other but otherwise entirely identical: parity conservation says (which also seems self-evident) that except for this mirror difference all subsequent evolution of these two systems should remain identical. This is precisely what the ^{60}Co experiment showed to be untrue. Now we know that parity violation exists nearly maximally in all weak processes (that is, those such as beta decay which involve the weak nuclear force). But before 1956, the 'self-evident' nature of parity made physicists unwilling even to entertain the possibility of its violation.

Discoveries in physics are made by looking in a new direction, quite often with the latest detection technology. The experiment on ^{60}Co was one such example. However, once parity was known to be violated, there were hundreds of other experiments that could be readily performed within a few days at many laboratories. This rare circumstance made the discovery of parity violation a particularly dramatic event.

At that time, almost the entire physics community wanted to participate in the overthrow of parity by undertaking experiments with other particles (pions, muons and hyperons), other nuclei and other decay processes. Wu, Ambler, Hayward, Hoppes and Hudson, and a few of their close colleagues who also engaged in the quest, chose to inform the world by holding a press conference at Columbia University on 15 January 1957, before the publication of their paper. This gave all



physicists an even start in verifying or extending this discovery. The human aspect of Wu's work on the experiment, her pleasure in collaborating with her co-authors at NBS and her joy in discovery are well documented in her moving account in *Adventures in Experimental Physics* (Gamma volume, World Science Education, Princeton, 1973).

Chien-Shiung Wu was born on 31 May 1912, shortly after the overthrow of the Qing (Manchu) Dynasty on 10 October of the preceding year. (The Chinese character for 'chien' means 'healthy' and for 'shiung' means 'male'.) Her birthplace was Liuhe, a small town near Shanghai. At that time the Western educational system did not exist in China and girls were educated at home, if at all. Through the encouragement of her father, she went to Shanghai to enrol in a class taught by Dr Hu Shi, the leading Chinese scholar (who revolutionized the Chinese language from the classical to the present form). Hu Shi immediately recognized her talents and gave her the confidence to enter the National Central University in Nanjing, where she received her BS degree in 1934.

Wu came to the United States in 1936 and received her PhD from the University of California, Berkeley, in 1940. She then taught at Smith College and Princeton University before joining Columbia in 1944, where she spent 37 years of her life doing research and teaching. She received many honours, including the Wolf prize (1976). Throughout her life, she was dedicated to science, meticulous in her experiments, and charming and gracious under all circumstances.

In 1934, Fermi proposed his theory of beta decay, but later on there were serious discrepancies between it and experiment. Then, in 1949 and 1950, through a series of experiments, Wu and her collaborators measured the allowed and forbidden beta spectra, corrected many previous mistakes,

disproved the alternative Konopinski-Uhlenbeck formulation and firmly established Fermi's theory. In 1963, she, Y. K. Lee and L. W. Mo observed the magnetic equivalent of the weak nuclear force, thereby confirming the symmetry between the weak and electromagnetic currents, which set the cornerstone for the later unification of these two basic forces into a single one — the electroweak force.

The ^{60}Co experiment of Wu *et al.* was a turning point in physics. The experiment itself established the violation of parity and charge conjugation symmetry, but it also started an avalanche of tests for all kinds of symmetry violations. In 1964, from the experiment by Christenson, Cronin, Fitch and Turlay, it was found that the joint symmetry CP and the time reversal symmetry T were both violated. However, as far as we know, the symmetry CPT remains intact; that is, we regain symmetry through the combined interchange of particle and antiparticle, right and left, and past and future.

The origin of these symmetry violations is still a mystery. Presumably, at the time of the Big Bang, the Universe was CP symmetric. Yet our very existence depends on CP asymmetry, which accounts for the preponderance of protons and electrons (matter) over antimatter in the Universe. CP violation has to have played a key role in making that happen, and thanks to the work of Wu and others an exciting new horizon opened up. The search for the origin and new signals of CP and T violations is the driving force of the high-energy B Facility accelerators in the United States and Japan.

When Madame Curie passed away, Einstein wrote as follows. All of his words apply equally to Chien-Shiung Wu. "At a time when a towering personality has come to the end of her life, let us not merely rest content with recalling what she has given to mankind in the fruits of her work. It is the moral qualities of its leading personalities that are perhaps of even greater significance for a generation and for the course of history than purely intellectual accomplishments. ... Her strength, her purity of will, ... her objectivity, her incorruptible judgement, all these were of a kind seldom found joined in a single individual. ... Once she had recognized a certain way as the right one, she pursued it without compromise and with extreme tenacity."

T. D. Lee

T. D. Lee is in the Department of Physics, Columbia University, 538 West 120th Street, New York, New York 10027, USA.

Rediscovery of the wild pig *Sus bucculentus*

Sus bucculentus, a species of wild pig from Indochina, was described more than a century ago¹ but has not been reported since and has remained in the 'mystery' category. We here report its rediscovery in the Annamite Range in Laos, an area which is becoming famous for the discovery of new and previously undescribed large mammals².

Sus bucculentus was described from two skulls¹ from Vietnam, and was claimed to differ from the much commoner *S. scrofa* (the widespread Eurasian wild pig from the same area) in skull characters and in the form of the lower canine of the male, in which it was said to resemble the warty pig (*S. verrucosus*) of Java. The whereabouts of the type specimens was unknown until recently. In July 1996, the type male skull was found by one of us (C. P. G.), unregistered but bearing an identification in Père Heude's handwriting, in the collection of

the Institute of Zoology, Academia Sinica, Beijing. It is inscribed "Bienhoa", an important colonial centre of the former French Indochina, and so doubtless a place of export rather than a locality. It does resemble *S. verrucosus* in many ways (deep pre-orbital fossa, elongated facial skeleton, high occipital crest, flat dorsal outline to the braincase and very long palate), but the canine is not like *S. verrucosus* as claimed in the original description¹. It has large premolars compared to the molars, in contrast with both *S. scrofa* and *S. verrucosus*; a primitive condition³.

A new specimen of *S. bucculentus*, a partial skull of a juvenile male, was obtained from indigenous hunters in the Annamite Range, Laos, in January 1995 by two of us (G. B. S. and K. K.; Fig. 1a). The locality of the new skull is Ban Ni Ghiang, 18° 19' N, 104° 44' E, a village on the Nam Gnouang. This can be taken as the first definite

locality for the species. The morphology is identical to that of the type skull (Fig. 1b) except for age differences.

Muscle tissue adhering to the new skull was used to extract and sequence a 327-base fragment of the 12S ribosomal mitochondrial gene, which was compared with material from a local wild *S. scrofa* (Fig. 1c) and a local domestic pig. The latter two were identical, but differed at 24 sites from the specimen of *S. bucculentus*. For a gene as conserved as the 12S ribosomal gene⁴, a difference as large as this is remarkable.

The rediscovery of this species is significant. It is another representative of a steadily enlarging clique of primitive large mammals whose ranges are more-or-less restricted to the Annamite Range², which lies along the central part of the Laos/Vietnam border. Here it lives alongside the much more widespread common wild pig, *S. scrofa*. This rugged region of igneous rocks, with localized karst cliffs and domes, still has large tracts of montane evergreen broad-leaved forest with an understorey of bamboo, palms and saplings, broken by swidden plots and regenerating forest.

The primitive status of *S. bucculentus* recalls two other recently described Annamite mammals: the bovid *Pseudoryx nghetinhensis*, which has been interpreted as alternatively a divergent member of the bovine (ox/buffalo/kudu) clade³ or the sister group of the caprine (sheep/goat/chamois) clade⁵; and the cervid *Megamuntiacus vuquangensis*, which is the sister group of the diverse genus *Muntiacus* (muntjac deer)³. These conclusions underline the significance of the Annamites as a biotically unique region where primitive taxa, long extinct elsewhere, have been able to survive into the late twentieth century.

Colin P. Groves

Department of Archaeology and Anthropology,
Australian National University,
Canberra, ACT 0200,
Australia

colin.groves@anu.edu.au

George B. Schaller

George Amato

Wildlife Conservation Society,
Bronx, New York 10460, USA

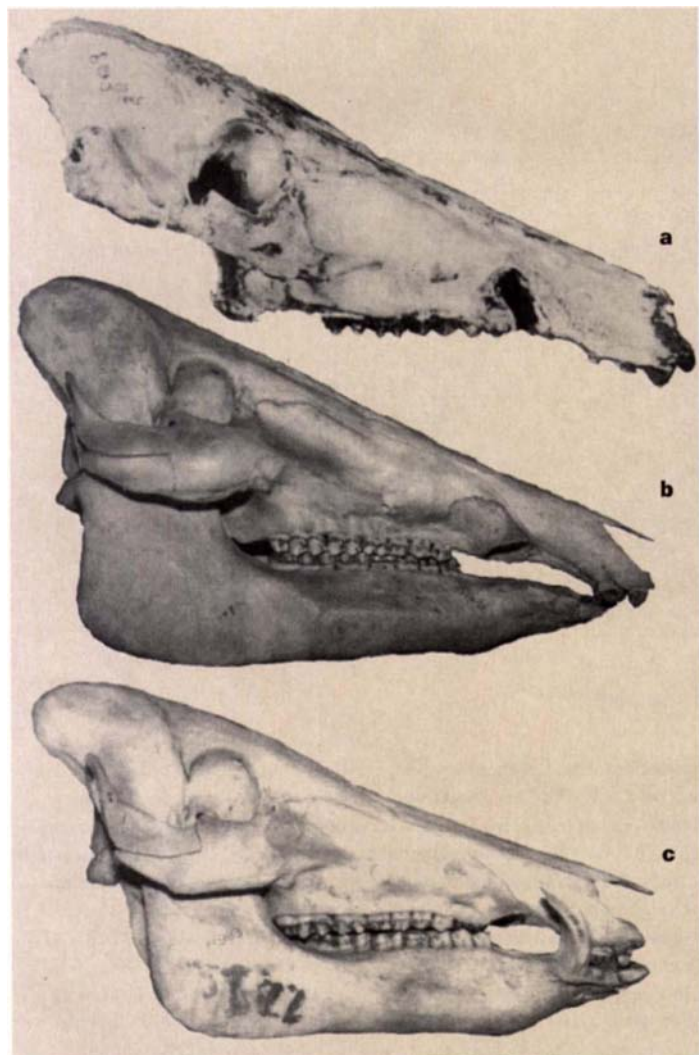
Khamkhoun Khounboline

Department of Forestry,
Vientiane,

Lao People's Democratic Republic

Figure 1 a,
Cranium of juvenile
Sus bucculentus,
with third molars
unerupted, from
Laos.

b, *S. bucculentus*
type skull, adult
male, labelled as
being from Bienhoa
(southern Vietnam).
c, *S. scrofa* adult male
from Indochina
(no exact locality).
Specimens are not to
scale: skull lengths
are respectively 354,
391 and 366 mm.
Distinctive visible
features are that
S. bucculentus has a
more elongated
facial skull than
S. scrofa; the
preorbital fossa is
deep and well
outlined; the dorsal
margin of the
braincase is straight,
not convex and the
occipital crest is
higher; and the
premolars are long
compared to the
molars.



1. Heude, P.-M. *Mémoires d'Histoire Naturelle de l'Empire Chinois*, 2, 85–115, 212–222; 4, 113–133 (Musée du Zikawel, Shanghai, 1892).
2. Schaller, G. B. & Rabinowitz, A. *Oryx* 29, 107–114 (1995).
3. Dung, V. V. et al. *Nature* 363, 443–445 (1993).
4. Amato, G. & Gatesy, J. in *Molecular Approaches to Ecology and Evolution* (eds Schierwater, B., Streit, B., Vagner, G. & DeSalle, R.) (Birkhauser, Basel 1994).
5. Thomas, H. *Mammalia* 58, 453–481 (1994).
6. Groves, C. P. & Dawson, S. *Mammalia* (in the press).

Bacterial DNA causes septic shock

Gram-positive bacteria lack lipopolysaccharide (LPS), an integral component of the outer membrane of Gram-negative bacteria, but their presence in tissues triggers inflammatory responses similar to LPS in terms of tumour-necrosis factor- α (TNF- α) and interleukin-1 production, cardiovascular abnormalities and death^{1,2}. To explain the responses it has been suggested that Gram-positive cell-wall components such as peptidoglycan or lipoteichoic acid are as toxic as LPS, yet experimental evidence is scarce³. Here we report a possible role of bacterial DNA in triggering septic shock during Gram-positive or Gram-negative bacterial infections.

Prokaryotic (bacterial) and eukaryotic DNA differs in structure. In eukaryotic DNA the dinucleotide motif 5'-CpG-3' is suppressed and mostly methylated^{4,5}. This structural difference seems to explain the finding that bacterial DNA and certain synthetic nucleotides containing unmethylated CpG motifs activate immune cells⁶⁻⁸. We have studied whether DNA prepared from Gram-positive or Gram-negative bacteria, or active synthetic oligonucleotides⁹ trigger rapid TNF- α production in macrophages and thus cause TNF- α -mediated toxic shock in D-galactosamine-sensitized mice.

DNA from Gram-positive or Gram-negative bacteria triggers fulminant TNF production in the macrophage cell line ANA-1 (Fig. 1a, b). The biological activity of Gram-positive DNA is DNase sensitive, whereas that of Gram-negative DNA is partly DNase resistant, presumably owing to trace contamination with LPS (Fig. 1b). Furthermore, the unmethylated synthetic oligonucleotide 1668 effectively triggered TNF- α release whereas methylation significantly reduced biological activity (Fig. 1c). At high concentrations of DNA, the magnitude of TNF responses induced were similar to that of LPS. At low concentrations LPS was more effective. Induction of TNF production by the various compounds in ANA-1 cells was preceded by nuclear translocation of the transcription factor NF- κ B and TNF- α messenger RNA expression. Our data may also corroborate a recent report that bacterial plasmid DNA induces TNF- α mRNA in macrophages¹⁰.

Mice injected intraperitoneally with Gram-positive DNA, Gram-negative DNA or active synthetic oligonucleotides showed high but transient serum concentrations of TNF- α , similar in kinetics and magnitude to those triggered by a bolus injection of LPS. Peritoneal macrophages from LPS non-responder C3H/HeJ mice produced TNF- α *in vitro* in response to bacterial

DNA and active oligonucleotide but not in response to LPS (Fig. 1d).

When mice were sensitized to TNF by D-galactosamine treatment¹¹, injection with Gram-positive DNA, Gram-negative DNA or the active oligonucleotide 1668 caused lethal toxic shock within 8–18 h. Toxic shock was due to acute liver failure mediated by TNF- α -triggered apoptotic cell death. Mice lacking the TNF receptor p55, or mice pre-injected with neutralizing anti-TNF- α monoclonal antibodies were unaffected. As SCID beige mice lacking B, T and NK cells also succumb to lethal shock after injection with bacterial DNA, we conclude that, *in vivo*, macrophages are the source of the TNF- α produced.

We have shown that bacterial DNA and active synthetic oligonucleotides cause rapid activation of macrophages *in vivo* and *in vitro*. Because of the high amounts of TNF- α produced, D-galactosamine-sensitized mice

succumb to lethal shock and our data may provide clues as to why Gram-positive bacteria can cause septic shock. Central to the biological significance of DNA-mediated immune-cell activation is the question of whether, at a natural site of bacterial infection, functional active bacterial DNA accumulates in sufficient amounts to cause cell activation. We anticipate that the threshold required is lowered if there are synergistic interactions between the signal pathway triggered by bacterial DNA and the CD14-aided pathway triggered by cell wall components.

**Tim Sparwasser, Thomas Miethke
Grayson Lipford, Katrin Borschert
Hans Häcker, Klaus Heeg
Hermann Wagner**

*Institute of Medical Microbiology, Immunology and Hygiene,
Technical University of Munich,
Trogerstrasse 9, D-81675 Munich, Germany*

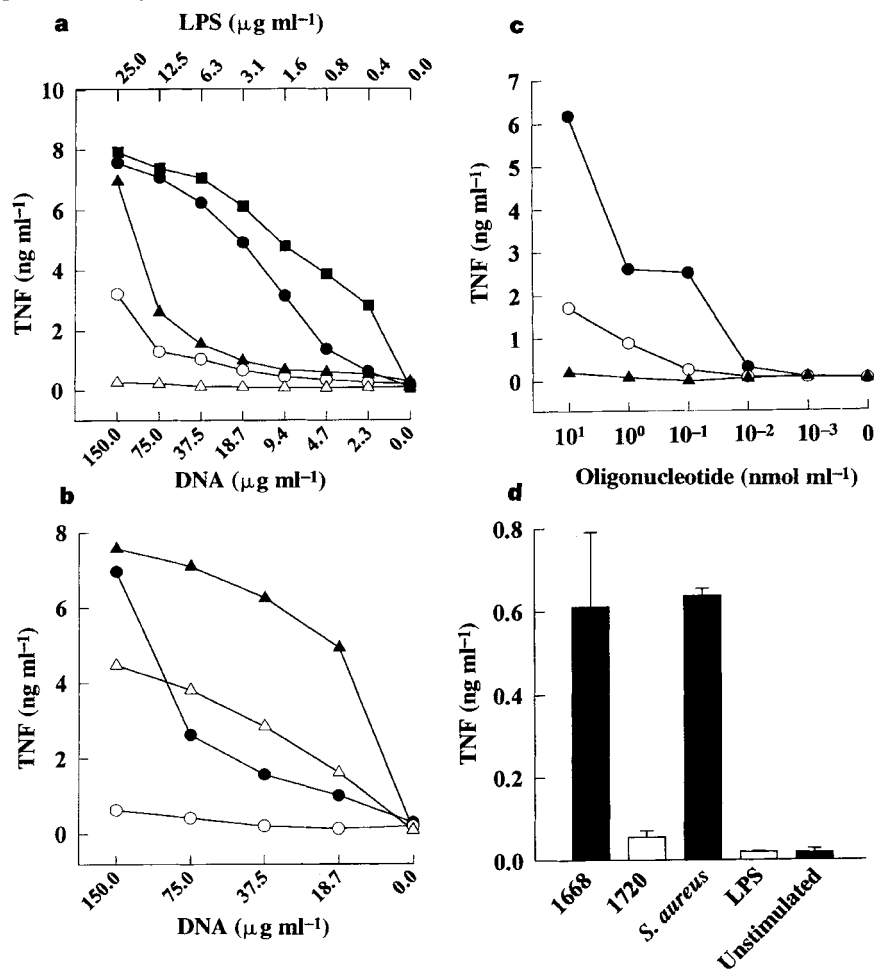


Figure 1 DNA or oligonucleotides stimulate the macrophage cell line ANA-1 (1×10^5 cells) to produce TNF- α . **a**, Titrated amounts of DNA from Gram-negative *Escherichia coli* (closed circles), Gram-positive *Streptococcus faecalis* (closed triangles), *Staphylococcus aureus* (open circles) or calf thymus (open triangles) were used and TNF- α release was measured by enzyme-linked immunosorbent assay (ELISA) after a 24-h culture period. LPS (closed squares) served as a control. **b**, DNA from *S. faecalis* (circles), or *E. coli* (triangles) was pre-treated with DNase (open symbols) or mock treated (closed symbols). **c**, ANA-1 cells were stimulated with oligonucleotide 1668 (5'-TCC ATG ACG TTC CTG ATG CT-3')⁹ (closed circles), methylated 1668 (open circles), or 1720 (5'-TCC ATG AGC TTC CTG ATG CT-3')⁹ (triangles) for 24 h. TNF- α levels were determined by ELISA. **d**, Peritoneal macrophages (1×10^6 per well) of LPS non-responder C3H/HeJ mice produce TNF- α on *in vitro* stimulation (24 h) with bacterial DNA ($100 \mu\text{g ml}^{-1}$) or active oligonucleotide (10 nmol ml^{-1}) but not with LPS ($10 \mu\text{g ml}^{-1}$).

1. Wakabayashi, G. *et al.* *J. Clin. Invest.* **87**, 1925–1935 (1991).
2. Hinshaw, L. B. *et al.* *J. Trauma* **33**, 568–573 (1992).
3. Kusunoki, T., Hailman, E., Juan, T. S., Lichenstein, H. S. & Wright, S. D. *J. Exp. Med.* **182**, 1673–1682 (1995).
4. Bird, A. P. *Nature* **321**, 209–213 (1986).
5. Sutter, D. & Doerfler, W. *Proc. Natl Acad. Sci. USA* **77**, 253–256 (1980).
6. Yamamoto, S. *et al.* *J. Immunol.* **148**, 4072–4076 (1992).
7. Pisetzky, D. S. *Immunity* **5**, 303–310 (1996).
8. Krieg, A. M., Yi, A. K. & Matson, S. *et al.* *Nature* **374**, 546–549 (1995).
9. Cowdery, J. S., Chace, J. H., Yi, A. K. & Krieg, A. M. *J. Immunol.* **156**, 4570–4575 (1996).
10. Stacey, K. J., Sweet, M. J. & Hume, D. A. *J. Immunol.* **157**, 2116–2122 (1996).
11. Miethke, T. *et al.* *J. Exp. Med.* **175**, 91–98 (1992).

Modern human origins backdated

Here we report datings of a hominid cranium (specimen KNM-ER 3884) and femur (KNM-ER 999) from the Lake Turkana region, Kenya, that indicate ages of around 270,000 and 300,000 years, respectively. These hominids might represent the oldest near-modern human specimens from anywhere in the world. Our datings and other recent evidence indicate that the chronological framework of *Homo sapiens* evolution in Africa needs to be revised.

The cranium from Ileret, northeast of Lake Turkana, was first reported in 1992 (ref. 1). It was found in deposits formerly attributed to the Guomde Formation², most of which has since been subsumed into the Chari Member of the Koobi Fora Formation. It was derived from undifferentiated later deposits, probably representing an age of 0.5 to 0.1 Myr (ref. 2). However, the specimen came from very close to the base of the latest Pleistocene/Holocene Galana Boi Formation¹ and so the hominid could be much younger.

We have now dated two different fragments of the cranium and a part of the femur, which derived from the same deposits³, by non-destructive γ -ray spectrometry. The technique has been used successfully to date other fossil hominids⁴. Activities of ^{234}U and ^{230}Th are determined from the γ -rays emitted at 53.3 KeV and 67.7 KeV, respectively⁵, by a high-purity germanium γ -ray detector (25% relative efficiency).

The two samples of the cranium yielded concordant U–Th ages of 272,000 years

(minimum age 159,000; indeterminate maximum age) and 279,000 years (minimum 162,000; maximum indeterminate). The U–Th date of 301,000 years (minimum 205,000, maximum indeterminate) for the femur supports the date of the cranium and indicates that both fossil hominid specimens came from very closely related stratigraphic levels. The U–Th dates are further supported by U–Pa dates of over 180,000 years for all three samples. Finally, an age of about 270,000 years for the cranium and 300,000 years for the femur are in agreement with the previous stratigraphic conclusions¹. The infinite upper errors leave open the possibility that the two hominids spanned a longer timescale, but they are both probably older than 180,000 years.

The cranium belongs to an adult individual. It consists of a large posterior part of the cranial vault including most of the occipital, parietals and temporals, a nearly complete supraorbital region and a maxillary part with all teeth. Preliminary estimation of endocranial capacity points to around 1,400 cm³. The posterior vault has thin walls (5–6 mm) and a lack of clear archaic features, and so shows close affinity to modern anatomy. In contrast, the torus-like supraorbitals are different from those seen in modern humans and closer to late archaic specimens like Florisbad and Laetoli H.18. Our observations indicate that the hominid might represent an archaic *Homo sapiens* or a transitional specimen very closely related to modern humans.

Comparative analyses showed that, in spite of its rather robust shaft, the femur has some modern features common among the earliest modern humans from Qafzeh and Skhul, Israel⁶. In view of the uranium-series dates, the femur might indicate that a very robust but basically modern morphology already existed in eastern Africa more than 200,000 years ago and probably as early as 300,000 years ago.

Such an early existence of near-modern transitional or late archaic *Homo sapiens* specimens, and the presence of early archaic *Homo sapiens* (Bodo, Ethiopia) at around 600,000 years ago⁷, as well as other recent datings of African archaic and early modern fossils (Eyasi, Florisbad, Singa)^{8–10} make a revision of the course of Middle Pleistocene

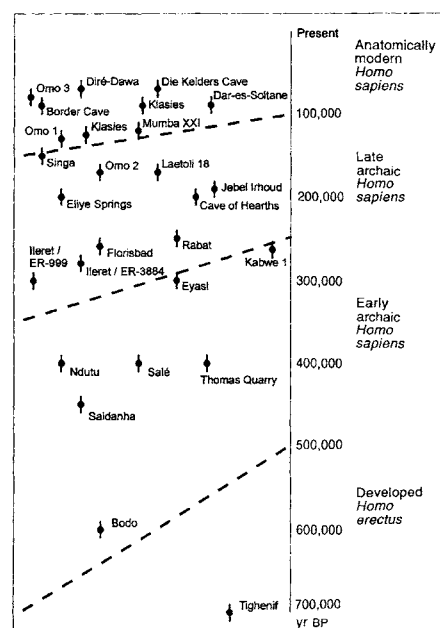


Figure 1 Revised scheme of *Homo sapiens* evolution in Africa. The transitional *Homo erectus*/archaic *Homo sapiens* period can now be dated to at least 700,000–500,000 years ago and the transition from early to late archaic *Homo sapiens* to around 350,000–250,000 years. The origin of modern *Homo sapiens* might go back to at least 150,000 years ago. The most likely ages of the specimens are given, but they have different qualities and errors.

evolution in Africa necessary (Fig. 1). Early and late archaic *Homo sapiens* and also the earliest modern humans seem to have existed considerably earlier than has been assumed¹¹. This revision further supports an early evolution towards modern humans in Africa and pushes the origin of archaic *Homo sapiens* back to the earliest Middle Pleistocene. This framework strongly affects the hypotheses of the emergence of archaic *Homo sapiens* outside Africa, for example the Ante-Neanderthals of Europe and archaic *Homo sapiens* of China.

Günter Bräuer

Institute of Human Biology,
University of Hamburg,
Allende-Platz 2, 20146 Hamburg,
Germany

Yuji Yokoyama, Christophe Falguères

Laboratoire de Préhistoire du Muséum
National d'Histoire Naturelle,
Institut de Paléontologie Humaine,
1 Rue René Panhard, 75013 Paris, France
Emma Mbua
National Museums of Kenya,
P.O. Box 40658, Nairobi, Kenya

Table 1 Results of uranium-series dating

	KNM-ER 999 (femur)	KNM-ER 3884 (cranium sample 1)	KNM-ER 3884 (sample 2)
Mass (g)	130.8	23.3	20.1
U p.p.m.	3.95	4.03	2.28
Ratio of nuclide activities			
$^{234}\text{U}/^{238}\text{U}$	2.253±0.256	3.773±0.760	4.343±0.757
$^{230}\text{Th}/^{234}\text{U}$	1.089±0.129	1.114±0.248	1.133±0.236
$^{230}\text{Th}/^{232}\text{Th}$	46	99	>200
$^{231}\text{Pa}/^{235}\text{U}$	1.083±0.060	1.185±0.178	1.309±0.210
Age (yr)			
U–Th (s.d.)	301,000 (+ ∞, –96,000)	272,000 (+ ∞, –113,000)	279,000 (+ ∞, –117,000)
U–Pa	>180,000	>180,000	>180,000

1. Bräuer, G., Leakey, R. E. & Mbua, E. in *Continuity or Replacement* (eds Bräuer, G. & Smith, F. H.) 111–119 (Balkema, Rotterdam, 1992).
2. Feibel, C. S., Brown, F. H. & McDougall, I. *Am. J. Physical Anthropol.* **78**, 595–622 (1989).
3. Leakey, R. E., Leakey, M. G. & Behrensmeier, A. K. in *Koobi Fora Research Project Vol. 1* (eds Leakey, M. G. & Leakey, R. E.) 86–182 (Oxford Univ. Press, 1978).

4. Yokoyama, Y., Falguères, C. & Bibron, R. in *L'Homme de Neandertal Vol. 1: La Chronologie* (ed. Otte, M.) 135–141 (Université de Liège, 1988).
5. Ivanovich, M., Latham, A. G. & Ku, T. L. in *Uranium-Series Disequilibrium: Applications to Earth, Marine and Environmental Sciences* (eds Ivanovich, M. & Harmon, R. S.), 62–94 (Clarendon, Oxford, 1992).
6. Trinkaus, E. *J. Hum. Evol.* **24**, 493–504 (1993).
7. Clark, J. D. *et al. Science* **264**, 1907–1910 (1994).
8. Grün, R. *et al. Nature* **382**, 500–501 (1996).
9. McDermott, F. *et al. J. Hum. Evol.* **31**, 507–516 (1996).
10. Bräuer, G. & Mabulla, A. in *Festschrift Jan Jelinek* (ed. Dočkalová, M.) (Anthropologie, Brno, in the press).
11. Bräuer, G. in *Continuity or Replacement* (eds Bräuer, G. & Smith, F. H.) 83–98 (Balkema, Rotterdam, 1992).

Endothelial nitric oxide synthase and LTP

Long-term potentiation (LTP) is often considered to be a cellular correlate of learning. During LTP induction in the CA1 region of the hippocampus, nitric oxide (NO) synthesized in the dendrites of pyramidal cells may carry retrograde signals from the postsynaptic to the presynaptic terminals^{1,2}. We show that LTP is defective in hippocampal slices from mice lacking functional endothelial nitric oxide synthase (eNOS) and from wild-type mice treated with a NOS inhibitor. The endothelial isoform of NOS seems to be required for the maintenance of LTP in the hippocampus.

Extracellular application of NOS inhibitors blocks LTP in the hippocampal CA1 area^{3–5}, although there are reports that LTP generated by strong stimuli is not sensitive to these drugs^{6–8}. Furthermore, in cultured hippocampal neurons, LTP is blocked by extracellular post- and presynaptic application of oxymyoglobin (which binds free NO), or by post-, but not presynaptic injection of a NOS inhibitor⁹. Exogenous NO paired with a weak, sub-threshold tetanic stimulus also induces LTP⁹.

Mice lacking a functional copy of the neuron-specific NOS isoform (nNOS) exhibit normal LTP, but LTP in nNOS knockouts is blocked by NOS inhibitors⁸. This may be explained by the finding that eNOS, initially believed to be present only in endothelial cells, is the main isoform in CA1 pyramidal cells¹⁰. Consequently, eNOS, not nNOS, may be responsible for synthesizing NO postsynaptically during LTP. Indeed, injecting hippocampal slices with an adenovirus vector containing a truncated, and hence not functional, eNOS gene (a putative dominant negative) blocks LTP at synapses in the CA1 stratum radiatum¹¹.

We inactivated the eNOS gene by replacing exons 24 and 25 with the neomycin-resistance gene in the embryonic stem cell line E14-1. Functional inactivation of eNOS was demonstrated by the lack of endothelial NO formation in eNOS^{-/-} mice, derived from two independently generated mutant

clones. The hippocampi of these eNOS-deficient animals had no obvious anatomical defects and apparently normal excitatory synaptic transmission in the CA1 region. Baseline test excitatory postsynaptic potential (EPSP) amplitude (30–40% of the maximum EPSP amplitude), was not significantly different for wild-type and eNOS^{-/-} mice (0.95 ± 0.05 mV for control slices, 1.08 ± 0.07 mV for eNOS^{-/-} slices, mean $\pm$ s.e.m.).

We chose a relatively weak LTP induction method (so as not to induce NO-independent LTP) that was highly sensitive to the NOS inhibitor *N*-nitro-*L*-arginine (NOARG). The method induced moderate potentiation in wild-type slices that lasted at least 90 min ($148.4 \pm 12.8\%$ 90 min after tetanus, $n = 11$). In the presence of NOARG, EPSPs exhibited short-term potentiation (STP) but not LTP, decaying gradually to baseline in less than 90 min ($94.9 \pm 4.2\%$ 90 min after tetanus, $n = 7$, Fig. 1a, b). After 90 min, the NOARG-treated

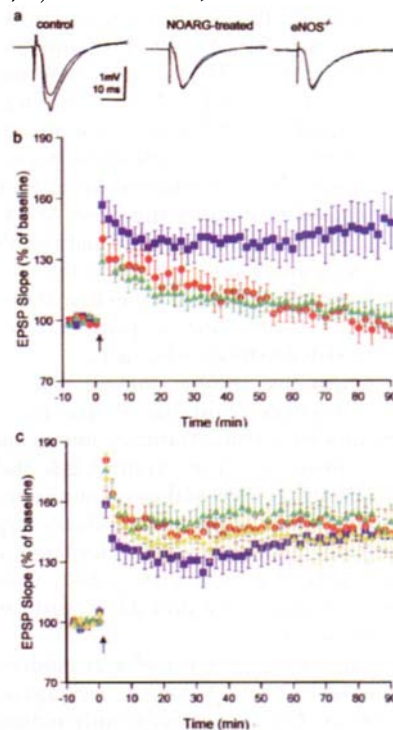


Figure 1 a, Superimposed field EPSPs, one recorded immediately before tetanus and one 90 min after tetanus, for wild-type slices, wild-type slices treated with 200 μM NOARG, and eNOS^{-/-} slices. b, Time course of mean normalized EPSP slopes in wild-type slices (blue squares), wild-type slices treated with 200 μM NOARG (red circles), and eNOS^{-/-} slices (green triangles) subjected to weak tetanic stimulation (3 trains of 10 pulses each, 100 Hz, 20 s inter-train interval, 40 μs pulses). c, Time course of mean normalized EPSP slopes in wild-type slices (blue squares), wild-type slices treated with 200 μM NOARG (red circles), eNOS^{-/-} slices (green triangles), and eNOS^{-/-} slices treated with 200 μM NOARG (yellow diamonds) subjected to strong tetanic stimulation (80 μs pulses). Tetanic stimulation is indicated by arrows. Vertical bars show s.e.m.

group was significantly different from controls ($P < 0.01$) but not significantly different from baseline. Using the same induction method, the eNOS^{-/-} mice exhibited STP but no LTP ($101.6 \pm 6.5\%$ 90 min after tetanus, $n = 10$, Fig. 1a, b). After 90 min, the potentiation level in the eNOS-deficient group was significantly different from wild-type slices ($P < 0.01$), and not significantly different from baseline.

Next, we examined LTP induced by a stronger tetanic stimulation. We found that doubling the pulse duration during tetanus only (from 40 to 80 μs) induced a level of LTP that was not significantly different in wild-type slices treated with NOARG compared to wild-type slices not treated with the drug ($143 \pm 11.8\%$ 90 min after tetanus for NOARG-treated slices, $n = 12$; $143 \pm 8.2\%$ for controls not treated with NOARG, $n = 8$, Fig. 1c). Using this tetanic stimulation, we observed robust LTP in eNOS^{-/-} slices ($150 \pm 8.6\%$, $n = 9$) that was not significantly different at 90 min from LTP in wild-type slices or in eNOS^{-/-} slices treated with NOARG ($142 \pm 13.3\%$, $n = 10$, Fig. 1c).

Our data are consistent with the hypothesis that NO synthesized by eNOS in postsynaptic CA1 pyramidal cells is a retrograde messenger required for the long-term maintenance, but not induction, of potentiation induced by weak stimuli. LTP induced by stronger stimuli does not seem to require NO, possibly because strong stimuli produce retrograde messengers in addition to NO. As LTP induced by strong stimuli in eNOS^{-/-} slices is not attenuated by NOARG, it seems unlikely that the LTP we observe is due to a compensatory activity of nNOS. The resemblance between our results for eNOS^{-/-} slices and for NOARG-treated wild-type slices suggests that eNOS is the main isoform participating in this process, and that any contribution from another isoform would be minimal.

Rachel I. Wilson, Ievgeniy Yanovsky

Axel Gödecke, David R. Stevens

Jürgen Schrader, Helmut L. Haas

Zentrum für Physiologie und Biomedizinisches

Forschungszentrum,

Heinrich-Heine-Universität,

D-40001 Düsseldorf, Germany

e-mail: haas@uni-duesseldorf.de

1. Garthwaite, J., Charles, S. & Chess-Williams, R. *Nature* **336**, 385–388 (1988).
2. Williams, J. H., Errington, M. L., Lynch, M. A. & Bliss, T. V. P. *Nature* **341**, 739–742 (1989).
3. Schuman, E. M. & Madison, D. V. *Science* **254**, 1503–1506 (1991).
4. O'Dell, T. J., Hawkins, R. D., Kandel, E. R. & Arancio, O. *Proc. Natl Acad. Sci. USA* **88**, 11285–11289 (1991).
5. Haley, J. E., Wilcox, G. L. & Chapman, P. F. *Neuron* **8**, 211–216 (1992).
6. Haley, J. E., Malen, P. L. & Chapman, P. F. *Neurosci. Lett.* **160**, 85–88 (1993).
7. Williams, J. H. *et al. Neuron* **11**, 877–884 (1993).
8. O'Dell, T. J. *et al. Science* **265**, 542–546 (1994).
9. Arancio, O. *et al. Cell* **87**, 1025–1035 (1996).
10. Dinerman, J. L., Dawson, T. M., Schell, M. J., Snowman, A. & Snyder, S. H. *Proc. Natl Acad. Sci. USA* **91**, 4214 (1994).
11. Kantor, D. B. *et al. Science* **274**, 1744–1748 (1996).

Human destinies and ultimate causes

Guns, Germs, and Steel: The Fates of Human Societies

by Jared Diamond

Norton: 1997. Pp. 480. \$27.50. Jonathan

Cape/Random House £18.99

Colin Renfrew

Why did the peoples of Europe come to dominate so much of the world? Or, in the words of the author's New Guinea friend Yali: "Why is it that you white people developed so much cargo [material goods] and brought it to New Guinea, but we black people had little cargo of our own?"

In setting out to answer this question and give the reasons for the regionally differing courses of history, Jared Diamond has written a book of remarkable scope: a history of the world in less than 500 pages which succeeds admirably, where so many others have failed, in analysing some of the basic workings of culture process. Its special quality is its clear-headed search for causes, for the underlying factors that gave some groups in human history the edge over others. Right at the start, the author rejects any ideas of genetic superiority — indeed he speaks of a "likely genetic disadvantage" of the Europeans — and he claims, with conviction, that "there really are broad patterns to history, and the search for their explanation is as productive as it is fascinating".

The book is a popular account and readable, leaving bibliographic references to a 'Further Reading' section at the end. The title is misleading, for the book focuses on a considerable time span, ranging over the past 10,000 years. Guns are little discussed and steel scarcely mentioned. But this should not deflect the reader from taking the work seriously, for it is one of the most coherent analyses of human history that I have read, and one that will stand on my bookshelf alongside Fernand Braudel's *The Mediterranean and the Mediterranean World in the Age of Philip II*. In its awareness of the underlying processes that have shaped human destinies, it is comparably encyclopaedic, yet broader in its geographical and chronological scope. In a manner comparable to Braudel's, it raises questions about the nature of historical explanation, and successfully avoids the methodological pitfalls that made Arnold Toynbee's *A Study of History* such an unsatisfactory contribution, for all its scope and erudition.

The author asserts: "History followed different courses for different peoples because of differences among peoples' environments, not because of biological differences among peoples themselves." Diamond is a physiologist, with experience in evolutionary biology and biogeography, and his focus is upon the environmentally related

Culture clash: the Spanish conquistador Francisco Pizarro who defeated the Incas in Peru.

diversification of human societies. Some will, therefore, view his approach as materialist, or even as determinist. Certainly his acknowledgement of the role of human individuals is a brief one, and his treatment inevitably lacks the fine texture of more detailed historical accounts — Braudel's *l'histoire événementielle* — yet many of his arguments carry conviction.

Diamond's earlier book *The Third Chimpanzee* (HarperCollins, New York, 1992) gives a very accessible account of human evolution and the present volume is, in a sense, a sequel, taking the year 11,000 BC as its starting-line. In Part I it dramatizes and exemplifies the initial question of the clash of cultures, with chapters on the successful domination by the Maori people of the Chatham Islands, and on the more famous encounter between the Spanish conquistador Pizarro and the Inca emperor Atahualpa. Part II devotes seven chapters (in effect the core of the book) to "The rise and spread of food production". Part III, in four chapters, proceeds with an examination of the evolution and significance of germs, writing, technology and systems of government. Part IV, "Around the world in five chapters", deals with the histories of Australia and New Guinea, of east Asia, of the Austronesian expansion, of Africa, and of Eurasia set in comparison with that of the Americas.

Diamond concludes with an epilogue, "The future of human history as a science", in which he asserts that the histories of human societies — like the histories of dinosaurs, nebulae and glaciers — may belong as much to the sciences as they do to the humanities.

Such a synthesis draws heavily on recent archaeological research, and the discussion of the origins and consequences of food pro-

duction in Part II constitutes the heart of the book. Diamond draws too on the work of the linguist Joseph Greenberg, whose classification of the African languages gives the key to the patterning observed on that continent. For his treatment of the Pacific and of South-East Asia, he relies on several archaeologists, notably Peter Bellwood, who have investigated the correlations between linguistics and archaeology.

Surprisingly, however, he makes little explicit reference (other than in the bibliography) to the impact that molecular genetics is now having upon our understanding of human dispersals. It may be that the rather easy, narrative style makes it difficult to refer directly to the underlying data (with which he is evidently familiar), but this informal approach does, on the other hand, allow him to bring in elements from his own personal experience, notably from his fieldwork in New Guinea. This includes first-hand acquaintance with hunter-gatherers and with simple agricultural communities (including that of his interlocutor Yali), revealing a personal understanding of, and sympathy with, contemporary non-urban societies and the people who live in them.

The great strength of the book lies in the assured manner in which the author takes the reader from what he terms the "proximate causes" of successful cultural dominance (the guns, germs and steel of the title) to what he regards as the "ultimate causes". These are, in effect, the basic geographical factors, not least the original distributions of potentially domesticable plants and animals, which allowed the early development and expansion of food production, notably in Eurasia. What he does not seek to confront,

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

and in effect avoids, is the whole series of controversies that underlie this apparently seamless synthesis. Greenberg's linguistic groupings have been criticized by many linguists, and this may be one reason why the author has very little to say about the Americas before Columbus.

I myself had one severe criticism of Diamond's *The Third Chimpanzee*, where I thought he misinterpreted the evidence for the domestication of the horse, setting the date of that domestication earlier than the current data allow, and confusing its early exploitation for food production around 4000 BC with the role of the mounted warrior in military conflict, which did not emerge until after 1500 BC and thus after the princely use of the horse-drawn chariot that is so clearly documented in the Near East and in the Mycenaean world. This may be one reason why his treatment of Europe (which is lumped together in a single chapter with the Americas) is sketchy. Another could be the current and very interesting disagreement among geneticists involving the notion of a significant population diffusion into Europe with the dispersal of farming (the 'demic diffusion' model of Cavalli-Sforza and his associates) being called into question by researchers at Oxford using data from mitochondrial DNA, only to be re-asserted by Pääbo and others who suggest that the effective mutation rate of the relevant mitochondrial DNA 'hot-spots' may have been gravely underestimated. (See L. L. Cavalli-Sforza, P. Menozzi and A. Piazza, *Science* 259, 639–646, 1993; M. Richards *et al.*, *Am. J. Hum. Genet.* 59, 185–203, 1996; and S. Pääbo, *Am. J. Hum. Genet.* 59, 493–496, 1996.)

Others may be critical of the limited role that is here expressly assigned to the operation of specifically human cognitive abilities. For there is little discussion about the differing natures of the cultures seen on the different continents: why Aztec culture took the form it did or why Chinese civilization differed from that of ancient Egypt.

There may, at first sight, be little in this book for the current school of 'post-processual' archaeologists, with their emphasis upon the uniqueness of each human society and their reluctance to make generalizations about human history. But to accept this would be to overlook the book's greatest strengths. It is willing to simplify and to generalize; and it *does* reach conclusions, about ultimate as well as proximate causes, that carry great conviction, and that have rarely, perhaps never, been stated so coherently and so effectively before. For that reason, and with few reservations, this book may be welcomed as one of the most important and readable works on the human past published in recent years. □

Colin Renfrew is director of the McDonald Institute for Archaeological Research, Downing Street, Cambridge CB2 3ER, UK.

Hitch-hiker's guide to the Solar System

The NASA Atlas of the Solar System
by Ronald Greeley and Raymond Batson
Cambridge University Press: 1997. Pp. 369.
\$150, £100

Carl D. Murray

By any criteria, the comprehensive exploration of our Solar System has to be regarded as one of the major scientific achievements of the century, if not the millennium. In the 40 years since the launch of Sputnik 1, a few generations have been privileged to witness the first images from the surface of the Moon, Mars and Venus, as well as *in situ* data returned from the robotic spacecraft that have visited all the planets from Mercury to Neptune.

For technical achievement and sheer audacity of purpose, the highlights of this exploration have to include the results from the two Voyager missions to the outer Solar System. In the 12 years after their launch in 1977, these NASA spacecraft have provided close-up images of Jupiter, Saturn, Uranus and Neptune and of their attendant moons and ring systems. Without such successful NASA missions, any attempt to produce a Solar System atlas would be as pointless as publishing a world atlas but restricting it to maps of Europe.

There is probably no ideal time to produce an atlas of the Solar System. With spacecraft such as Mars Pathfinder and Mars Global Surveyor about to provide another order-of-magnitude increase in our knowledge of one planet, and with the Cassini-Huygens mission to Saturn waiting

in the wings, a delay of another decade or so would produce a more complete work. However, the natural pause produced by the end of the radar mapping of Venus by Magellan, and the full analysis of the data from the Galileo spacecraft orbiting Jupiter, means that now is as good a time as any to undertake this ambitious project.

The understated goal is "to provide a set of maps of uniform format and consistent scales, organised by planetary system for all of the objects seen thus far". This is a massive undertaking and, thanks to the map-making skills of the US Geological Survey, the results are truly spectacular.

Although the authors make no claim to have produced a textbook, the atlas is much more than a collection of maps. As well as a general introduction, background information is provided for each planetary system, complete with colourful, informative diagrams and clear explanations. Every solid surface that has been viewed with sufficient resolution has its corresponding geological, reference and shaded relief map, with the occasional bonus of a colour photo mosaic for objects such as Mars, Io, Europa and Triton. For reasons of clarity, not all features are named on all maps, but the atlas includes a complete gazetteer listing every named feature from Aananin (a crater on the Saturnian moon Rhea) to Zwicky (a lunar crater), together with its latitude and longitude on the respective body and a brief description of the name's origin.

This is a reference book *par excellence*. The design is superb, with the uniform format making it easy to carry out comparative planetology. Small touches such as the outlines of planetary orbits at the top of the page are particularly helpful, with the orbit of the current system highlighted for the browser



The martian volcano Olympus Mons, photographed by Viking Orbiter.

— the atlas's equivalent of 'you are here'.

Some features do not work so well. There is a valiant attempt to use a series of colourful strips with different scales to illustrate the locations of the wide range of orbits within a given planetary system, but the result is confusing. There are also the inevitable minor errors and inconsistencies associated with such an enormous enterprise. The most glaring mistake is that the relief map of the Neptunian moon Triton is labelled "Titania", the name of a much smaller and less interesting Uranian satellite; perhaps we can be generous and attribute this to the action of some mischievous spirit.

The sheer size of Greeley and Batson's finished product makes it difficult to lift, thereby reinforcing the impression that this is a reference work. There are inevitable comparisons with the lavish world atlases from previous centuries. Even the frequent blank areas on maps of satellites and planets — regions not imaged during brief spacecraft fly-bys — are reminiscent of the labelling by map-makers of large tracts of land in Earth's Southern Hemisphere as "Terra Incognita". While some might feel cheated by paying good money for empty space, such regions are a sobering reminder of our level of ignorance and the exciting task of exploration that lies ahead.

In a few decades, the blanks will be filled in by the next generation of explorers, both robot and human. Until then, this atlas will serve as the definitive guide to the Solar System and a fitting testimony to NASA's achievements in this field. □

Carl Murray is in the Astronomy Unit, Queen Mary and Westfield College, University of London, London E1 4NS, UK.

Custom-made beast of burden

The Nature of Horses: Exploring Equine Evolution, Intelligence and Behavior

by Stephen Budiansky
Free Press: 1997. Pp. 261. \$30

Horses Through Time

Edited by Sandra L. Olsen
Roberts Rinehart: 1996. Pp. 222. \$35.
Distributed in the UK by Airlift Books, £25

Marian Stamp Dawkins

It is definitely an exaggeration to claim, as the blurb to Stephen Budiansky's *The Nature of Horses* does, that the book contains "everything horse lovers, novice riders, and professional trainers alike need to know about one of our most beloved and fascinating animals". There are no instructions about how to train horses, for example, or what to feed them on, but there are nevertheless some fascinating facts and ideas. One

Not a horse chestnut, but *Mares by an Oak Tree* by George Stubbs.

of the most memorable is Budiansky's view that the horse, being a large mammal vulnerable to climate change, would have become extinct at the end of the Ice Age as many other species did, were it not for domestication by humans. North America, which had been the cradle of equid evolution for 55 million years, had no horses by 10,000 years ago; they were reintroduced by Christopher Columbus in 1494. The American Indians, who had never seen the animals before but subsequently of course made famous use of them, called them 'big dogs'.

The horse as we know it was probably domesticated in the Ukraine about 6,000 years ago. Budiansky argues that it was brought back, quite literally, from the brink of extinction by happy accidents of history that produced an animal preadapted for human use. The horse has a social system with rules of dominance and subordination that makes it particularly tractable for obeying humans; it is capable of carrying heavy burdens over long distances; it has a fortunate lack of horns or antlers; and it even possesses a convenient gap in its cheek teeth where humans can place a bit. Instead of becoming extinct, the animal made an explosive comeback by 4,000 years ago, so horse remains are common in archaeological sites from Asia to Britain.

The book contains some informative, if not particularly original, chapters on the animal's sensory system, and useful descriptions of gaits and the way the horse manages to be so large yet so fast. The horse-racing fraternity comes in for some criticism for its lack of understanding of genetics. The Jockey Club in America, for example, refuses to register horses that have been bred through artificial insemination or embryo transplants. Racing thoroughbreds are inbred anyway, so it is hardly surprising that attempts at breeding horses to race faster have not been very successful. In fact, speed records for the animals have remained virtually unchanged for the past 50 years, despite meticulous attention to stud books

and breeding. Budiansky recommends opening up horse-racing to all breeds, with the prediction that cross-bred horses would soon be leaving the thoroughbreds behind.

While Budiansky's book is well-written, full of useful information and certainly enjoyable, I found myself vaguely annoyed by two things.

One was the point already mentioned: that the book simply did not live up to its claim to cover everything one might want to know about horses. I would have thought, for example, that any book that dealt with human-horse relationships should have offered more on the different ways of training horses, from the time-honoured ways of subduing them to the newer ways which involve establishing a rapport with a social equal. Some of the claims that have been made for these new methods are spectacular (the Queen is said to be impressed with them). But do they work?

The second was that much of his writing draws upon the work of other people, and it is sometimes difficult to know whether the ideas are his own or not. This concern is particularly obvious if one reads Budiansky's book, as I did, soon after reading *Horses Through Time*, edited by Sandra Olsen. Budiansky not only draws heavily on the work of several of the authors in the latter book but even uses several of the same photographs, such as one of an Assyrian frieze and a particularly dramatic one of a horse-skin suspended on a pole.

Olsen's book has an altogether different flavour, even though it covers much of the same ground and is concerned with the evolution and domestication of the horse as well as its continuing relationship with humans. It contains ten different essays by nine different authors, giving it the inevitably slightly disjointed feel from which any multi-authored volume is liable to suffer.

There are some excellent and well illustrated chapters, including an authoritative description by Richard Hulbert of the horse's ancestry, and others on what is

known about how the horse was domesticated (surprisingly little). Juliet Clutton-Brock contributes an uplifting chapter on the horse in history, reminding us how important horses have been. Without the horse, Alexander the Great and Genghis Khan would probably not have been so successful, and the Spaniards could not have destroyed the Incas. The Roman Empire also depended largely on horses. Perhaps, she suggests, without raiding armies on horseback, many ancient civilizations would still be flourishing. At the very least, human history would have been different without the horse to accelerate the movement of people and their cultures and languages around the world.

The choice of subjects for the book, while interesting, does seem a little arbitrary. Why include a section on advanced techniques in horse medicine that have existed for only a short time, for example, but say very little about the animal's behaviour, which has been around for much longer?

Perhaps I am asking too much of books that both set out to describe the long history of human interaction with the horse and would not, despite their respective claims, have space to be truly comprehensive. But I could not help feeling disappointed that, although both described horses and their relationship to us in a fascinating way, neither addressed what is probably the hottest topic of all: the ethics of that relationship and what we do to horses. □

Marian Stamp Dawkins is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

A plague of weapons

The Eleventh Plague: The Politics of Biological and Chemical Warfare

by Leonard A. Cole

W.H. Freeman: 1996. Pp. 284. \$22.95, £24.95

Alastair Hay

Leonard Cole ranks chemical and biological warfare as the "eleventh plague". The first ten, described in the book of *Exodus*, were reputedly inflicted by God on the Egyptians for their failure to release the Jews from slavery. The arsenal available to the Almighty would appear to have been more limited than it is now for, say, the United States or Iraq. Today, the ubiquity of the chemical industry and the widespread use of fermentation processes offer endless opportunities for synthesizing chemicals in bulk or culturing pathogenic organisms.

If the technology is no barrier to producing agents of chemical and biological warfare, what are the constraints on their use? This is the principal question addressed by Cole in



Art transplant

The surgeon Sir Roy Calne performed the first liver transplant in Europe. But he is an artist too, and has painted transplant patients and fellow surgeons. Many of these paintings are reproduced in his book *Art, Surgery and*

Transplantation (William & Wilkins Europe, £45), along with other paintings about medicine by better-known artists. One such is *The Sick Child* by Edvard Munch (shown here), which hangs in the Tate Gallery in London.

this critical, well researched and eminently readable book. After reviewing the strengths and weaknesses of the various treaties proscribing chemical and biological warfare — deterrence and the huge costs required to maintain an adequate defence programme — Cole concludes, perhaps unsurprisingly, that a range of factors are required to deter use. These include verifiable treaties that allow inspections; defence programmes (justifiable both in cost and to the public they are meant to protect); good intelligence; and a much more outspoken position on the immorality of these weapons.

Three treaties prohibit the use of agents of chemical and biological warfare. Two of these, the 1925 Geneva Protocol and the 1972 Biological Weapons Convention, have no policing provisions and no sanctions. Apart from the exhortation to forswear the use of these agents, the treaties can do no more than create a climate hostile to their use. Cole points out that Iraq, even though a party to both treaties, developed a significant chemical and biological warfare capability, with the result that thousands of Iranians and Kurds were killed by Iraqi mustard and nerve gases.

This legacy has resulted in policing and inspection becoming two of the cornerstones of the 1993 Chemical Weapons Convention, which will become binding international law later this year. Many view it as the key to controlling chemical warfare, but two of the key signatories have still not ratified the treaty in their own parliaments and so are not bound by its provisions. The United States, which has more than 30,000 tonnes of chemical warfare

agents, and Russia, which has nearly 40,000 tonnes, are still outside the fold. Both countries were major participants in all the negotiations that led to the treaty; it is embarrassing for both, and very disquieting for everybody else, that they are not signed-up members.

The control of biological weapons will be influenced by the success of the convention. There is no doubt that a policing system is urgently required for biological warfare agents. As for a defence programme on biological warfare, Cole observes that, in the United States, "hundreds of millions of dollars have been spent" on it, and yet, in his opinion, the "US citizenry is no better protected from a biological attack now than when the army's research program began half a century ago". Cole believes any future improvements in defence will be only marginal.

Vast resources have been spent building chemical arsenals; far more will be required to dismantle them. Research on biological warfare agents has been less costly but significant in the United States, the former Soviet Union and Iraq. Information about activities in other countries is more scanty.

There is still time to rein in these activities and to persuade countries that their security is enhanced by being party to treaties that prohibit chemical and biological warfare. In agreeing to abide by these treaties, countries are also taking the moral high ground. Chemical and biological warfare is immoral, and there are many who would class it in the way Cole does, as a disease. □

Alastair Hay is in the School of Medicine, University of Leeds, Leeds LS2 9JT, UK.

Die either on the gallows, or of the pox

The Great Pox: The French Disease in Renaissance Europe

by John Arrizabalaga, John Henderson and Roger French

Yale University Press: 1997. Pp. 352. \$35, £25

William Bynum

When AIDS forced itself into the public's consciousness in the early 1980s, historians were encouraged to find parallels between our modern scourge and the sudden appearance of syphilis in Europe half a millennium before. Ever willing to oblige, my colleagues and I pointed out a few similarities: the panic generated by a new disease, the search for scapegoats, the moralizing — even the conclusion that both afflictions were the 'wages of sin'.

It is one of the many virtues of this admirable book to demonstrate how generic, if not actually specious, such parallels are. AIDS rates only one passing mention; nor is the authors' subject the spirochaetal disease that we know as syphilis. Rather, Arrizabalaga, Henderson and French seek to examine what the condition commonly called the 'French disease', or *mal francese*, meant to doctors, laymen and sufferers from the late fifteenth century. Their concern is with perceptions and responses rather than the prehistory of a modern disease. There are no new insights here on that hardy perennial: "Was syphilis a New World import brought back by Columbus' sailors, or an Old World disease newly transmuted and, perhaps, newly transmitted sexually?"

That this disease, which broke out in Naples, Florence, Ferrara and other Italian cities from 1496, did appear to many contemporaries to be new is not in doubt. And by calling it the French disease, they implicated the invading troops of King Charles VIII of France in its spread. The attempt of the French to call it the Neapolitan disease was less successful. The Columbus hypothesis was not part of the original debate, however; nor did everyone subscribe to the proposition that this was truly a new scourge. Nicolò Leonicensi (1428–1524) — the long-lived and tireless advocate of humanistic medical reform through a return to the solidity of Hippocrates, Galen and other Greek authors — argued, with the author of *Ecclesiastes*, that there is no new thing under the sun. Hippocrates had seen it all and, under other names such as leprosy or elephantiasis, the French disease had ever plagued mankind.

Nor, throughout the sixteenth century, was the French disease ever exclusively

venereal in its perceived mode of transmission. For one thing, priests, bishops, cardinals and popes suffered alongside the common people. 'Virtuous ladies' as well as prostitutes and soldiers fell victim to it. For another, the French disease did not begin, as does modern syphilis, with a genital chancre, but exhibited a mass of symptoms and signs. These included rash (the great pox, as distinguished from smallpox), intense pain in the bones and joints, sores, scabs and pustules. The genitals were implicated often enough to place unclean sexual intercourse among the contributing factors, however, and *lues venerea* became one of the disease's many names during the century.

The sexual dimension certainly reinforced the religious and moral overtones that the disease acquired early on, but in the religious context of Renaissance Europe, the existence of disease itself posed the very questions that had long ago plagued Job. The French disease was thought particularly loathsome, however (in Strasbourg, even the lepers shunned those suffering from it), and Jews, prostitutes and other marginal groups were sometimes singled out for castigation.

Nevertheless, from the extensive evidence of this volume, it is a mistake to isolate the French disease from more general medical concerns of the age. From the early series of public disputations on its nature and causes — the wonderfully named doctors Pistorius and Pollich who disputed in Germany could be characters from *Cymbeline* or another late-Shakespearean romance — to the continuing debates about the best form of treatment, professional medical issues loom large. Mercury and guaiac bark emerged as the principal drug alternatives. Guaiac's New World origin eventually gave credence to the Columbus hypothesis (God placed remedies near diseases), but, in the sixteenth century, physicians were more concerned with justifying their preference by appealing to the relationship between the properties of drugs and the nature of the disease.

Mercury was a traditional treatment for a variety of skin disorders, but it was disagreeable and could be dangerous. It was a mainstay of surgeons, Paracelsians and quacks; learned physicians, who recognized its potency, insisted that its real danger lay in incorrect administration by incompetent hands. By contrast, decoctions of guaiac wood were new and easier on the patient, but also more expensive and, according to many, less effective. Despite its cost (the Fugger family in Germany made a fortune

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

French made: a patient suffers from the pox in this woodcut from 1520.

importing it), guaiac was commonly used in Italian hospitals for 'incurables'.

These hospitals, here analysed in detail, began to be founded in many Italian cities from the early sixteenth century as a response to the new disease. Records from the Spedale di San Giacomo in Rome belie the common historical caricature of early hospitals as 'gateways to death'; mortality rates there were in the range of 10 to 15 per cent. Admissions were predominantly male, although females had higher mortality rates, probably because they were likely to be more badly nourished and more desperately ill than their male counterparts. The clinical descriptions do not permit modern diagnoses, but many patients were perceived at the time as suffering from the French disease. When the 'holy wood' (as guaiac was known) was being administered, patients were especially keen to be admitted.

Arrizabalaga and his colleagues have elucidated these and many other aspects of their subject. They argue convincingly that the French disease catalysed both institutional and conceptual change within Renaissance medicine. They are chary of calling this change 'progress'. Rather, they insist that our forebears had knowledge that was just as meaningful to them as our own is to us. Or, as the radical medical reformer Paracelsus (1493–1541) once remarked: "What was true for the Greeks is not true for us. Truth must be born in its own land." And, he might have added, in its own time. □

William Bynum is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.

The hunting of the helix

The Double Helix

by James D. Watson

Weidenfeld & Nicholson: 1997. Pp. 174. £20

The Eighth Day of Creation

by Horace Freeland Judson

Cold Spring Harbor Laboratory Press: 1996.

Pp. 693. \$55 (hbk), \$39 (pbk)

Walter Gratzer

When it was first published in 1968, everyone in the business seized on *The Double Helix* with eager anticipation: we wanted not so much to have our minds improved by the science as to be entertained by some of Jim Watson's grosser anecdotes. Nor were we disappointed, and the reviews by Watson's friends and former friends propelled his book into *The New York Times*' best-sellers list.

Watson was probably surprised by the tumult, for he had not, I believe, envisaged a *succès de scandale*; he meant the book, rather, to stand as a serious contribution to history by a protagonist in one of the great intellectual adventures of the century. Of course, as Oscar Wilde observed, autobiography offers a unique opportunity to tell the truth about other people, and it is no secret that Watson's fellow Nobel laureates, Francis Crick and Maurice Wilkins, made strenuous efforts to prevent publication of what they saw as a parody of the events. But Crick has since told his own version of the story, as also now have others — many of them to Horace Judson — so one can take one's choice.

One of Crick's objections to Watson's vastly entertaining portrayal is that it debases a glorious achievement by turning it into a mere race to Stockholm against contenders in Pasadena and London. In Crick's recollection, science was the only spur. Yet it is undeniable that one of the most painful experiences that can overtake a scientist is to have someone else shoot the bird he has in his sights. Judson suggests, for instance, that this is largely what explains Erwin Chargaff's ill-natured reaction to Watson and Crick's discovery and to all that has flowed from it: the problem that he cherished was suddenly and brutally snatched away from him — and, worse yet, by a pair of amateur arrivistes.

What makes *The Eighth Day of Creation* so absorbing, besides of course Judson's skill as a narrator, is the sense of a voyage into unknown regions. When Einstein conceived the Special Theory of Relativity, he wrote to his mother that he had discovered not just a new country but a new continent. So it was when the principles of molecular genetics and control and the structure of proteins started to fall into place. Bliss was it in that dawn to be alive.

The personalities, the dramas, the anecdotes are all here — Crick; George Gamow

Watson and Crick: protagonists in a great intellectual adventure.

(Geo., known to himself and his friends as Joe) and his tie club (membership limited to 20, one for each amino acid); François Jacob's first encounter with the Olympian André Lwoff; Sydney Brenner's excited dance on the beach in California when he and his friends suddenly perceived how the messenger might be snared; and many another good story.

For Brenner, Seymour Benzer and Jacob, all leaders in the quest, much of the allure went out of molecular biology when it moved from the periphery to the centre and became a discipline, with departments in universities, textbooks and degree courses. Brenner laments that today it is all brute force and little ingenuity, and that the conviviality has vanished with the increase in numbers of those joining in. The pioneers, who wrote few papers and discovered so much, have been succeeded by what used to be known to the press corps as *pisseurs de copie*.

A striking feature of the history that Judson relates is the extent to which the new science was dominated by a coterie of friends in a half-dozen or so laboratories. They kept in close contact (without the benefit of e-mail or fax), exchanged results and ideas, met to talk and planned experiments. They displayed most often a lordly disdain for outsiders, and they received at least one nasty shock when Marshall Nirenberg — then an unknown in an unfashionable lab — found the key to the coding problem. When rumours began to circulate, the members of the club reassured each other that there was no need to take notice; and when Nirenberg reported, at the Moscow meeting in 1961, that polyuridylic acid was a messenger for polyphenylalanine synthesis, few were present at the session. Alfred Tissières, then with Watson at Harvard, reminisced that they had all been furious with themselves for not trying Nirenberg's experiment first. It was put

about — and Judson nails the *canard* — that Nirenberg's was an accidental discovery, the polyuridylic acid assay a mere control.

All this is recounted in Judson's original edition, and the new material in this expanded version does not amount to much. It includes some observations by Jacob on the character of his difficult colleague Jacques Monod, now safely dead (and theatrical to the end, with some lofty, probably well-rehearsed, last words on his lips), and there is an enlarged description of Frederick Sanger's early protein sequences and their impact on thinking about the code. For the rest, the text is essentially unchanged and Max Perutz, for instance, is still "today in his mid-sixties" (he is actually now in his eighties, with his intellectual, and indeed literary, powers undimmed).

Substantial additions are two appendices, which reassess the contributions of Rosalind Franklin and Chargaff. Their histories show that premature death can be a shrewd career move, for Franklin became almost instantly a feminist numen. Watson's rather unchivalrous treatment of her in his book inflamed her admirers; and a regrettable biography published in 1975 made strident assertions that Franklin had been a lone woman in a den of male bigots at King's College, and that her work had been devalued and her character traduced.

Judson has gone to some lengths to disperse the mist of obfuscation. The laboratory at King's, he found, housed 31 research workers, of whom eight were women. He traced and communicated with all but two, and also the honorary biological adviser to the laboratory, the redoubtable Dame Honor Fell. Their testimony is unanimous: there was no discrimination; the atmosphere was informal; and the women, together with some of their male colleagues but never the unclub-

bable Franklin, made it a congenial custom to lunch together at the Strand Palace Hotel of a Saturday.

What must assuredly have soured her life at King's and, more importantly, her relations with Wilkins, was the written guarantee that J.T. Randall, the director of the laboratory (not, by the way, a "Scottish physicist", as Judson has him), had given her that the DNA problem would be hers alone and that Wilkins would be pursuing other interests; meanwhile, Randall had given Wilkins to understand that the new recruit would assist him in the crystallography of DNA. The mischief that this caused cannot be doubted, but neither can Franklin's unaccommodating character and hostility towards the colleague with whom she should have shared her ideas.

When I came to the laboratory at King's, Franklin was already dead, but she was not remembered with affection. Before a laboratory seminar on DNA, she circulated, as Judson relates, black-edged cards, announcing the demise of the helix and concluding: "It is hoped that Dr M. H. F. Wilkins will speak in memory of the late helix." A witness of these events told me that when Franklin discovered an error in a calculation of a Patterson function, which had apparently misled her, she begged him not to tell Wilkins. Franklin's statue as a scientist remains unchallenged, but Judson takes the view that attempts at canonization have done a disservice to her memory.

Chargaff is another whose work on DNA has sometimes been seen as undervalued. He was one of a brilliant and cantankerous *galère* of Central Europeans given shelter by Columbia University in the late 1930s. Grateful or not, they found it difficult to conceal their scorn for their American colleagues, who spoke no Latin or Greek, and they detested the insidious incursion of the 'hard sell' into science.

Chargaff, moreover, could no more retain a wounding witticism in his mouth than a hot potato. To be sure, his description of Crick as "a fading racing tout" and something out of Hogarth departs from the norms of polite academic discourse, and Judson properly describes it as embarrassing. The Watson and Crick of 1952 were neither of them exactly Chargaff's type. Wilkins, by contrast, remembers him as nothing other than kind and helpful. It is certain only that, had Chargaff merely kept his mouth shut, instead of denouncing Watson and Crick and their hateful model, his place in the pantheon would have been a deal more secure. One might mention in this regard the late Lord Todd, a man of impregnable self-esteem, who was irked by talk of Watson and Crick's structure. They had not, he would declare, determined the structure of DNA; he, Todd, had determined its structure, they only its conformation.

Judson's book and Watson's have been reissued, not perhaps for the last time, as

classics. *The Double Helix* appears with the original foreword of saintly forbearance by Sir Lawrence Bragg, and a new introduction by Steve Jones which reveals that even the most accomplished of performers can have an off-day.

A historian has mused that the memory of man is too frail a thread on which to hang history; Judson's achievement, in drawing out the memories of so many participants in the epic of molecular biology and weaving them into a single robust skein, is magisterial. His work fittingly commemorates a golden age which already seems as remote as that of Darwin and Huxley. □

Walter Gratzer is at The Randall Institute, King's College, London, 26-29 Drury Lane, London WC2B 5RL, UK.

The solution to pollution?

Laboratory Earth: The Planetary Gamble We Can't Afford To Lose

by Stephen Schneider

Basic Books: 1997. Pp.174. \$20. Weidenfeld & Nicholson £11.99

John Houghton

Humans have rarely anticipated environmental pollution; action and pollution control have generally followed on from environmental disasters or unacceptable pollution damage. However, the potential magnitude of damage to the environment that humans are now able to cause, and the speed with which such damage can occur, are such that it is now necessary for the impact to be assessed and action taken before the damage occurs rather than afterwards. This is particularly true with pollution that is global in its extent, such as that leading to climate change as a result of the increased burning of fossil fuels or tropical deforestation.

Because of this need to anticipate, questions are bound to be asked about the accuracy of scientific arguments, and the extent to which the increased pollution is likely to matter. Answers can be found only by viewing the problem in a wide context, taking account of both the science and the politics of the problem. The science of climate change is highly interdisciplinary, involving all of the main disciplines of natural and social science. Because few can claim expertise in more than a small number of the appropriate disciplines, any adequate assessment of the problem necessarily involves a large community of scientists.

It is in this wide context of science and policy that Stephen Schneider addresses the issue of global warming and climate change. The likely changes projected by climate models are compared with changes that have occurred over the millions of years of the

Earth's climatic history; the modelling of these past changes has helped to validate the climate models used for future projections.

In presenting the effects of likely climate change, Schneider does not focus on the impact on the resources required by humans, such as water resources; a more balanced account would have given these critical effects on humans more attention. Rather, he homes in strongly on the loss of biodiversity, devoting a whole chapter to it — despite the large uncertainties that tend to fog the issue — arguing that, because biodiversity is fundamentally irreplaceable, its loss must be considered as the most serious effect of rapid climate change.

The latter part of the book is devoted to the question of how policy decisions should be approached in such a complex field. Here the book has a distinctive North American flavour; there has been much more of an open battle in the United States between some of the scientists and the lobbyists working for the energy industries. This in turn has stimulated a lively debate between ecologists, who are trying to preserve the planet, industrialists, who think any problem can be fixed *post hoc*, and economists, who are trying to gauge the likely effect on the sacred cow of economic growth.

For logical decision-making, 'integrated assessment' sounds an appropriate formula, but is it realistic to suppose that all the relevant factors can be weighed by any of our existing technical machinery? Of course, cost-benefit analysis can be attempted, but the things that we really value are hard if not impossible to turn into numbers, especially numbers representing money. Schneider emphasizes the value of going through the discipline and arguments that might lead to 'integrated assessment', but sensibly draws back from presenting it as the final arbiter. In what will be for many readers the most interesting part of the book, he attempts to explain the much more complex relationship between policy judgements and the scientific, technological and economic arguments (with all their uncertainties) that assist in the formulation of policy and the prescription of action.

Schneider has done much to raise awareness of global environmental issues, particularly through his contributions to the media. He speaks fast and lucidly, and his writing possesses much of the same racy style. The book is laced with considerable technical detail that is informative to those with a training in science or policy, but even readers who are not scientifically trained will gain a broad understanding of the scope and scale of one of the most important environmental challenges of our day. □

Sir John Houghton is chairman of the Royal Commission on Environmental Pollution, Church House, Great Smith Street, London SW1P 3BZ, UK.

A shrinking discipline

A History of Psychiatry: From the Era of the Asylum to the Age of Prozac

by Edward Shorter

Wiley: 1997. Pp. 436. \$30, £19.99

John C. Marshall

Psychiatry has always been the most controversial of medical disciplines. So controversial, indeed, that many of its more eclectic practitioners have not thought it a discipline, while others have denied that it formed part of medicine. Edward Shorter, professor in the history of medicine at the University of Toronto, has now written a polemical history of psychiatry that should prove as contentious as the practice thereof.

From the word go, Shorter makes his own position clear: "If there is one central intellectual reality at the end of the twentieth century, it is that the biological approach to psychiatry — treating mental illness as a genetically influenced disorder of brain chemistry — has been a smashing [*sic*] success." Shorter distinguishes *absolutely* between 'biological psychiatry' and what he calls the "biopsychosocial model of illness". These perspectives, he claims, "are polar opposites, in that both cannot be true at the same time". He continues: "Either one's depression is due to a biologically influenced imbalance in one's neurotransmitters, perhaps activated by stress, or it stems from some psychodynamic process in one's unconscious mind." Words (almost) fail one: does Shorter really believe that psychological states are disjoint from brain states?

Thankfully, Shorter's bizarre philosophy does not inhibit him from some forthright judgements on 'physical' treatments, including drug therapies. The reader needs a strong stomach to cope with Shorter's descriptions of caring physicians rendering their patients comatose with insulin; provoking convulsions with Metrazol and fever with malarial injections; and committing all other manner of grievous bodily harm. Here, for example, is Dr Hatcher describing how he performs transcortical lobotomies: "Nothing to it. I take a sort of medical icepick, hold it like this, bop it through the bones just above the eyeball, push it up into the brain, swiggle it around, cut the brain fibres like this, and that's it. The patient doesn't feel a thing."

By comparison, what Shorter calls "the psychoanalytic hiatus" sounds relatively benign. One does, nevertheless, wonder what point Shorter thinks he is making when he claims that "[i]t was above all among the middle-class Jews of Berlin, Budapest and Vienna that psychoanalysis proved such a

hit". Or when he glibly asserts that "It seems to be the case that Jews overconsume most psychiatric services in proportion to their numbers in the population".

Shorter is equally plain-spoken (and equally ambiguous) on the pharmacological industry. In a section entitled "Maintaining Market Share", he argues that such 'conditions' as youthful exuberance, exposure to the news on television, and plain ordinary unhappiness have been medicalized into attention-deficit disorder, post-traumatic stress disorder and mild depression, to keep the psychiatrists in work and the drug companies in profit. Yet Shorter remains convinced that there is such a thing as "real psychiatric illness", uncontaminated by the fact that "psychiatrists have an obvious self-interest in pathologizing human behaviour". These 'real' pathologies include bipolar psychosis (manic depression) and schizophrenia. And, for such biological diseases, the patient is enjoined to swallow the pill of the month, despite the fact that the prescribing physician knows little of what links the chemistry to the symptomatology.

Shorter seems unperturbed by this lacuna, for he has foreseen a new problem for the profession. Could the pharmacological victory be Pyrrhic? By their very success, the chemical conquistadors may have done themselves out of a job: "Every time a psychiatric disorder became medicalized, it disappeared from psychiatry", a vexation that Shorter exemplifies with the loss of "neurosyphilis to the internists, mental retardation to the paediatricians, and stroke to the neurologists".

Is all then lost? Apparently not. The average consultation time in psychiatry is purportedly 40 minutes (significantly longer than in most specialties); time enough, it would seem, for the psychiatrist

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

The insane at Bethlehem Hospital, London. From *The Rake's Progress* by William Hogarth.

to talk to the patient. "This combination of psychotherapy plus medication," Shorter claims, "represents the most effective of all approaches in dealing with disorders of the brain and mind." And why can this "psychotherapeutic side" *not* be "hived off to the psychologists and social workers, who were more intensively trained as therapists"? Simple: "The history of medicine suggests that patients derive some kind of bonus from the knowledge that they are dealing with a physician"! "Catharsis", claims Shorter, is heightened when "telling one's story" to a doctor.

The Freudians — who Shorter despises, and whose ideas, he asserts, "are now vanishing like the last snows of winter" — should thus find instant gratification in Shorter's history (but only, of course, if they are medically qualified). Everyone else would be well advised to take a few tranquillizers (or better still, a stiff whisky) before turning to the preface. There, Shorter informs us that it is merely a "trendy" notion among intellectuals that psychiatrists might certify "those who otherwise would be challenging the established order". Many of my friends from the former Soviet Union will be amused to read that this does "not correspond to what actually happened".

Even more disturbing is Shorter's confidence that "[p]sychiatry is, to be sure, the ultimate rulemaker of acceptable behaviour through its ability to specify what counts as 'crazy'". Funny, I always thought that the 'ultimate rulemaker' (on Earth) was the highest court of the land. As the good soldier Švejk argued, the doctors had no right to throw him out of the lunatic asylum without first giving him lunch. □

John C. Marshall is at the Neuropsychology Unit, University Department of Clinical Neurology, Radcliffe Infirmary, Woodstock Road, Oxford OX2 6HE, UK.

Engineering just-so stories

Invention by Design: How Engineers get from Thought to Thing

by Henry Petroski

Harvard University Press: 1996. Pp. 242.

\$24.95, £16.50

David Jones

The keypad of a telephone has the digits 1,2,3 at the top. That of a calculator has the digits 7,8,9 in this position. Most of us use both devices every day. How serious is this failure of design?

This is one of the questions left as an exercise for the reader by Henry Petroski in his survey of the design history of a selection of familiar objects. He starts small, with paper-clips, pencils, zip fasteners, beverage cans and fax machines, and finishes big, with aircraft, water distribution systems, bridges and buildings. His themes are the Darwinian, evolutionary nature of the engineering design process and the slow, often irregular optimization of its artefacts. The book is a series of stories about how his chosen products came to be the way they are.

His vision, like that of all good Darwinians, is optimistic. Bad designs fail in the market-place while good ones thrive, so technology steadily improves. Here is the 'chemisealed' pencil of 1930, with its graphite centre bonded to its wooden surround. It thus forms a superior cantilever beam, less likely to break at the tip. Here is the aluminium beverage can, in its triumphant progress towards ever-lower weight, easier opening and (with the invention of the retained opening tab) more efficient recycling. Here is the fax machine, its steady improvement driven appropriately by the Japanese, whose ideographic script is almost impossible to send by teleprinter.

The technical aspects of these stories are

very appealing. To see a pencil as a cantilever beam or a beverage can as a pressure vessel is to feel the power of engineering insight. The commercial aspects also have their subtlety. In the long history of the paper-clip, innumerable patents were taken out on various designs, and almost all of them failed. The now-universal 'Gem' clip, which dates from about 1890, was never patented. Its most successful exponent, William Middlebrook of Connecticut, patented instead a machine for making it. It could churn out Gem clips so cheaply that they came to dominate the market.

But beneath the technicalities, engineering design is a social activity. Its consistent problem is the conveying of facts and ideas from one mind to another. Even the old-style inventor in his garret or garage faces this problem. He may do all the thinking, designing and constructing himself, but he still has to sell his vision to his backers. If he succeeds, his vision will reach the market-place, where his customers may (or may not) accept it as well.

Modern engineering, especially that of large systems such as bridges and aircraft, is beyond any single inventor. It needs teams of designers. The difficulties of exchanging ideas within a team are serious: they might be expected to rise as the square of the number of members.

The traditional solution is that of hierarchy. Nevil Shute in *No Highway* (1948) paints a striking portrait of the ferocious aircraft designer E. P. Prendergast, "...a man of immensely powerful will, capable of imposing his idea and his way of doing things on each of a hundred draughtsmen, so that each one of them is too terrified to insert any of his own ideas".

Prendergast was based on the great

Barnes Wallis. His dogmatic approach contrasts strongly with Petroski's exposition of the democratic, computerized system used by Boeing half a century later to design the 777 aircraft. Designs and drawings for the whole craft were stored in eight of IBM's largest mainframe computers, and each of thousands of designers could call down any drawing to his workstation. If two drawings conflicted, the clash was immediately apparent. Petroski does not say if the conflicts were settled by democratic vote or authoritarian diktat.

But the Boeing model is still highly unusual. One traditional mental divide is still very much with us: that between designers and engineers. The designers dream up a component or even a complete product; they then (as Petroski puts it) "throw it over the wall" to the wretched engineers, who have to figure out some way of making it. If the engineers succeed in making the thing, and the sales force in selling it widely, the designers will soon be back at work on a new model — usually without even bothering to discover what was wrong with the old one.

Some manufacturers, it is said, deliberately rush a flawed product to market, and allow the resulting customer complaints to guide the design of its replacement. But even this rudimentary feedback is often ignored. The most crucial mental gulf of all, and the most seldom bridged, is that between designers and users. I would have welcomed more discussion of this problem; its neglect or misunderstanding leads to numerous bad designs.

In the electronic age, the deadliest temptation for the designer is to add more features. Petroski's account of the fax machine is one of steady improvement, standardization, and simplification. Yet many electronically controlled goods — such as video cassette recorders, washing machines and photocopiers — are almost unusably complicated.

Pride of place in the dysfunctional design stakes must go to computer software. The most amazing achievement of the computer software industry is its continuing cancellation of the steady and staggering gains made by the computer hardware industry. Analysts look in vain for any serious increase in overall productivity from the vast global investment in computers over the past 20 years. Let us hope that, even now, Darwin is grooming some unsuspected software mutation for the mighty and much-needed kill.

As for the problem of the incompatible keypads, it cannot be long before somebody makes a telephone with calculator attachment, or possibly a calculator with telephone attachment. We should then at least discover which of the numeric arrangements is more fitted to survive. □

David Jones is in the Chemistry Department at the University of Newcastle, Newcastle upon Tyne NE1 7RU, UK.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Fly by wire: a computerized system allowed thousands of designers to work on the Boeing 777.

Deadly mushrooms

Feux Follets et Champignons Nucléaires

by Georges Charpak and Richard L. Garwin
Editions Odile Jacob: 1997. Pp. 382. FF150

Hubert Curien

Will-o'-the-wisps, the natural manifestation of the combustion of gases emitted by the dead rotting in their cemeteries, used to strike unreasoned terror into the hearts of our forebears. This dread finds psychological resonance today in the blind fear of radiation. The title of this book, *Will-o'-the-wisps and Nuclear Mushrooms*, is doubly intriguing, but just two paragraphs are devoted to jack-o'-lanterns; the remainder concentrates on nuclear radiation.

The authors, one French, the other American, are well-known physicists. Georges Charpak works in Geneva at the European Laboratory for Particle Physics (CERN) and in Paris, and won the Nobel prize for physics in 1992 for the design and construction of particularly efficient particle detectors. Richard L. Garwin devoted much of his talent to developing US nuclear weapons and security systems during the Cold War. President Bill Clinton has recently called on him to collaborate with Russian colleagues in reviewing strategy for the control and destruction of plutonium stocks. Charpak and Garwin therefore constitute a first-rate fount of knowledge of the civil and military applications of nuclear physics; they are a well-tuned duo in which each player is a fine soloist.

The authors begin with clear explanations of the principle of mass-energy equivalence and of the properties of neutrons.

These provide a sound basis for reflection and for the formulation of recommendations. The authors give a level-headed account, but indulge in appealing whimsy by presenting chain reactions in nuclear reactors in the form of a tragi-comedy, wittily illustrated by Sempé, the author of several collections of cartoons appreciated by young and old alike.

Having learned almost all there is to know about the neutron, the reader is introduced to the production of energy by nuclear fission using water-cooled and fast-neutron reactors. There are plenty of statistics, tables and graphs, and a demonstration that the pollution generated by the burning of fossil fuels is unquestionably greater than that caused by nuclear energy. But it is not so much pollution as the fear of accidents that is psychologically problematic.

The comparative study of the accidents at Chernobyl and Three Mile Island is revealing. In one case, human losses were great; in the other, the damage was obviously considerable but essentially financial. In both cases, human error played an important, if not major, role, and the behaviour of the "nucleocrats" scarcely helped. Was it not a Soviet official who opined after the Chernobyl disaster that "Science requires victims"?

These accidents naturally led to a resurgence of the demonizing of nuclear energy, but also, and more rationally, to renewed demands for the accurate evaluation of the effects of radiation on living matter. The granite on which our houses are built is radioactive; so too is the human body. Is there a safety threshold? And if so, where should it be set? Charpak and Garwin objectively set out the research and conclusions of those for and against. Even when the data are not filtered by those of one or other persua-

sion, considerable scope for uncertainty lingers. But we should not be tempted to salve our consciences by arguing irresponsibly that, in any case, the harm done by smoking may well exceed that caused by radiation.

Nor should we sidestep our duty to act wisely over the management of radioactive nuclear waste. No other industry has ever had to consider the consequences of its activities over a period of thousands, if not millions, of years, even if some doubtless reacted when these consequences seemed likely to impinge appreciably on natural resources or alter irreversibly the environmental balance.

The authors consider that, in the distant future, highly advanced fast-neutron reactors could provide a rational solution to the problem of waste. They also believe that new approaches to the use of energy from nuclear fission, such as that described by Carlo Rubbia with his "Energy Amplifier", are worthy of serious consideration. Rubbia proposes the use of a subcritical reactor, thereby excluding all possibility of an uncontrolled chain reaction. This reactor would be coupled to a proton accelerator in which the protons collide with lead nuclei and generate the neutrons required to trigger fission. Such a reactor would consume thorium, which is naturally more abundant than uranium, and could also use plutonium and the radioactive waste produced by other types of nuclear power stations.

"One hundred years! Give the scientists one hundred years", urge Charpak and Garwin. Scientists are now confronted by a vast field of endeavour covering optimal storage or destruction of waste, and viable alternative methods of energy production. Let us hope that these one hundred years run their course unthwarted by a global catastrophe precipitated by inept management of the Cold War legacy of some 60,000 atomic bombs!

And what of the military applications? What are the likely effects of a one-megaton bomb exploded at a height of 2,000 metres? Soviet experts considered that 600 such weapons would have been sufficient to destroy 80% of the industrial output of the United States. But the number of bombs was rationalized in terms of uncertainty about how many would penetrate the enemy's defences. Still fresh in our memories is the Strategic Defense Initiative, more popularly known as "Star Wars". Would it really have protected the United States? It certainly gave a powerful fillip to the electronics industries and, sadly, to the arms race. This point is made by the reproduction here of a particularly clear-sighted and germane article first published in 1984 by H. Bethe, R. Garwin, K. Gottfried and H. Kendall.

Can nuclear weapons be eliminated altogether? The world has entered an active phase of nuclear disarmament, but some military men and politicians still mistrust

Artists have felines too



The naturalist Daniel Giraud Eliot commissioned the German artist Joseph Wolf to illustrate his *A Monograph of the Felidae*, first published in 1883. All 43 plates, including this one of Geoffroy's cat (*Felis geoffroyi*, now *Oncofelis geoffroyi*), are part of a collection at the American Museum of Natural History. They have been reproduced in full in *The Family of Cats* (Pomegranate, \$16.95).

simple dissuasion, clinging instead to the notion of absolute protection through nuclear weapons. Charnak and Garwin propose a form of disarmament in which most nuclear warheads are rapidly scrapped and the components put into storage, while a minimal number of weapons are primed for launching. But how small is minimal?

In considering the advisability, even the necessity, of nuclear testing, slight differences emerge in the American and European viewpoints. The authors concur on specific proposals for a massive and rapid reduction in stockpiles of nuclear weapons. The United States and Russia should cut their respective arsenals to 2,000 warheads over the next two years, and the three other self-confessed nuclear powers would start scaling down their own nuclear weaponry once the American and Russian stocks descend below 1,000. A worldwide treaty could then be negotiated for the complete elimination of nuclear weapons.

The authors take as the title to their conclusion "Scientists, politicians, utopians, realists, ecologists: a common purpose". They may have cast their net wide, but who would criticize them?

Charnak and Garwin have written an excellent book in which they clearly separate comment and interpretation from information. This allows readers to learn without losing the freedom to make judgements for themselves, guided in their decisions by two trustworthy advocates. □

Hubert Curien is the president of the *Academia Europaea*, and is emeritus professor of crystallography at the University of Paris, Paris 75252, France.

Gladiator for science

Huxley: Evolution's High Priest

by Adrian Desmond

Michael Joseph: 1997. Pp. 340. £20

Crispin Tickell

This is the second part of Adrian Desmond's biography of Thomas Henry Huxley, and covers the period between 1870 and his death in 1895. It follows, rather awkwardly, from the first part (published in 1994 and reviewed in *Nature* 375, 300–301; 1994) as if the reader had just put the latter down and was still familiar with the cast of characters. This second part draws upon material from London's Imperial College that was not available to previous biographers.

During his life, and afterwards, Huxley was given many labels: 'Darwin's bulldog', 'Pope Huxley', 'Huxley Eikonoklastes', 'the devil's disciple' and now 'evolution's high priest'. In my view, these are misleading in different ways. During the great changes in nineteenth-century thinking about the natural world, Huxley was more of a partner to

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

'Darwin's bulldog': T. H. Huxley, painted by Alphonse Legros.

Darwin than his bulldog or disciple. He reproached himself for not having spotted the processes of natural selection, but his particular and crowning achievements lay elsewhere. More than anyone, he established the central role of science in a society living in an intellectual world remote from our own. Together, he and Darwin shifted the paradigm, in what was then the leading country of the world, in a fashion that can be compared with Newton's achievements almost 200 years before. Huxley did for science what others have since tried to do for the environment. The statues of Huxley and Darwin stand benevolently together in London's Natural History Museum.

Huxley was a man of amazing parts. Born in a relatively poor dissenting family, he learned how to infiltrate the ruling establishment by speaking its language and understanding its ways. Establishments rarely mind radicals if they can present themselves wearing best bib and tucker. Huxley became a member of the Athenaeum, president of the Royal Society and, eventually, a privy counsellor. In some respects, he was a conservative (with a small 'c'), and suspicious of socialism and the political revolutionaries of his time.

This respectability helped him to carry out an expert demolition job on some religious beliefs, and particularly on the Church itself. Just before his death he wrote: "I am not afraid of the priests in the long run. Scientific method is the white ant which will slowly but surely destroy their fortifications." But he became something of a priest himself, not so much of evolution (as suggested in the subtitle of this book), but of agnosticism (his own word) and of an approach encapsulated in a letter he wrote in 1886: "I am too much of a sceptic to deny the possibility of anything." Thus he long held

back, at least in public, from unqualified endorsement of evolution by natural selection, and described theories as "excellent servants but very bad masters".

Not least, Huxley was a creative scientist, combining detailed practical work in the laboratory with the rare gift of exposition of general ideas. Looking back, it is fascinating how far he saw ahead, predicting, for example, the role of living organisms in shaping their physical environment (now part of Gaia theory); the way that *Penicillium* mould stops bacterial growth; the likelihood that birds descended from dinosaurs; and the doctrine of garbage in, garbage out. As a prolific intellectual bruiser, he was without peer in his time. Woe betide those who wandered into his sights: from Bishop Wilberforce in 1860 to the ageing Gladstone in the 1880s, Huxley dealt with them in crisp and, above all, decisive fashion. He also enjoyed wide popularity. For the young H.G. Wells, he was a continuing inspiration. On Huxley's death, a dock labourer, living, as he wrote, from hand to mouth, sent a shilling postal order as a contribution to his memorial.

The role Huxley played was achieved at some personal cost. He was subject to depression and always felt driven. Until near the end, he was usually short of money. Although he was the patriarch of a happy extended family with a much-loved wife, his eldest son died in infancy, his siblings suffered a variety of misfortunes and the most talented of his daughters suffered from mental instability and died young. Any charges of immorality aimed at his thinking did not extend to his way of life. Even so, his eldest granddaughter remembered being ostracized at school for her descent from 'the great atheist', and he was the object of abuse from zealots who wrote during his last illness to rejoice in his coming prospects in hell.

Adrian Desmond brings together new information about Huxley and presents it in lively, albeit breathless, fashion. I found the best and most illuminating part occurred towards the end of the book, when Desmond seeks to place his subject in historical perspective. For the life itself, he chooses to write according to a "ciné theory of narration... a highly mediated celluloid construct, where the camera pans across a stage". I do not think this method works very well, and I hope it does not become a model for others. Too often, the result is a cross between *Time* magazine and television soundbites, in which continuity and coherence are lost. Clichés pile up: "the old world scornfully eyed the new"; people "chortle", "crow", "chuckle"; and so on. The book as a whole is a stock cube in need of hot water to make it digestible.

Yet it brings out facts about Huxley that are new, and there are insights of real value. It is the product of painstaking research and will itself become part of the archive. Huxley

once admitted that mildness in one of his articles was not "the softness of the answer which turneth away wrath, but... that of the pillow which smothered Desdemona". As the foremost gladiator of science in his day, he knew how to use every weapon. He thereby changed the mind of his own generation and of our own as well. □

Sir Crispin Tickell is at Green College, University of Oxford, Woodstock Road, Oxford OX2 6HG, UK.

Roots theories and racist interpretations

Race and Human Evolution: A Fatal Attraction

by Milford Wolpoff and Rachel Caspari
Simon & Schuster: 1996. Pp. 462. \$25, £17.99.

Leslie C. Aiello

This book attempts to answer one of the most fundamental questions of human existence: why are we here? In doing this, the authors put forward one of the best and most clearly articulated cases so far for the multi-regional theory of modern human origins. Perhaps more importantly, they couch their description and defence of this theory in the context of the intermeshed history of human origins studies, racism and politics. For this reason alone, students of human evolution should read and take heed of this book. It is too easy for us to sit in our universities and ignore the intellectual history of the theories that we put forward and the ways in which they may be manipulated by the political and cultural contexts in which we work.

Wolpoff and Caspari are careful to define multiregionalism as a theory that suggests that for some two million years human populations have been entwined in a network of widespread peoples who evolved together because they met and interbred, giving the races today many ancestors, not a single common one. This gloss on multiregionalism is different from that commonly associated with the theory by the popular press and many who argue against the idea.

The popular interpretation of multiregionalism is that races of *Homo erectus* in Africa, Asia and Europe gave rise to the modern human races living in those areas of the world. The implication is that, during the long course of evolution, isolation rather than continuity was the rule. Wolpoff and Caspari argue strongly that this unjustified emphasis on isolation in human evolution exposes the theory of multiregionalism to equally unjustified racist interpretations. At the same time, they stress that it also renders the theory unreasonable in the context of Darwinian evolution. Throughout the book, they return again and again to the point that modern multiregionalism is not a polygenetic theory of racial evolution.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

A *Homo erectus* skull: racism rears its ugly head in interpretations of human origins.

Polygenism is the largely nineteenth-century idea that human races have different origins and separate histories. Most polygenetic theories formulated after Darwin posited some type of orthogenetic impetus driving the evolution of these isolated races in the same general direction. The more separate the human races are, the easier it is to argue that some humans are more human than others, and to build a justification for colonialism, racism and slavery. This is precisely how these theories were used.

Wolpoff and Caspari show that, although polygenism is a convenient prerequisite to racism, all polygenists were not necessarily racists, and that all adherents of the alternative view, monogenism, were not necessarily enlightened egalitarians. Rather, racism then as well as now is governed by a complicated interplay of influences from natural history and taxonomic thinking, religion, politics and, from the middle of the nineteenth century, Darwinism in its various guises. Much of the core of the book traces these various threads through the personalities involved in late nineteenth and early twentieth century physical anthropology and racial studies. In so doing, it establishes intellectual genealogies which make fascinating reading for anyone familiar with the work of the personalities involved.

This discussion sets the stage for the defence of one of the most respected hominid anatomists of the twentieth century, Franz Weidenreich, and his perhaps less respected polycentric theory of human origins. It is no accident that the book builds up to this climax, because Weidenreich's polycentric theory is the immediate intellectual forerunner of multiregionalism. Wolpoff and Caspari argue strongly that Weidenreich viewed human races as continuous, interconnected and ephemeral entities with considerable time depth. To their mind, poly-

centrism has been seriously misrepresented by modern palaeoanthropology as simply another polygenetic theory, incompatible with modern evolutionary biology and open to racist interpretations.

By association, they also feel that multiregionalism is unjustifiably tainted. Wolpoff and Caspari are also careful to distance multiregionalism from another and perhaps infamous version of polycentrism put forward in the 1960s by Carlton Coon. In emphasizing the discontinuous nature of human races, Coon's theory harks back to some of the worst excesses of nineteenth-century academic racism.

On this basis, the last three chapters of the book are devoted to the defence of modern multiregionalism. Most of the arguments presented are well known to palaeoanthropologists, but the clarity with which they are presented makes *Race and Human Evolution* a welcome contribution to the literature. However, the book is not in itself going to convince the majority of the academic community that alternative single-origins theories for the evolution of modern humans are dead. This is particularly the case because Wolpoff and Caspari insist on polarizing the debate. They argue against what they insist on calling the "Eve theory", the most extreme version of a variety of single-origins models for modern human evolution.

The real question is whether their own theory of multiregionalism is genetically viable. Is it really possible that modern human features could have arisen individually and at different times and places throughout the vast geographical range occupied by Pleistocene human populations, and coalesced into modern humans and modern human diversity as we know it today? Many palaeoanthropologists and evolutionary biologists will remain to be convinced even after reading this book. We can expect this lively debate to continue for a long while to come. □

Leslie C. Aiello is in the Department of Anthropology, University College London, Gower Street, London WC1E 6BT, UK.

ADVERTISEMENTS

NEW!

Walter G. Neumann

Universe and Brain

The world beyond the limits
of physics and philosophy

german-english • ISBN 3-925679-42-1
7.96 £ • 13.54 \$ • 19.90 DM

edition scientia
Raabestraße 18 • D-30177 Hannover

READER ENQUIRY NO. 549

Nitride-based semiconductors for blue and green light-emitting devices

F. A. Ponce & D. P. Bour

Recent advances in fabrication technologies for the semiconducting nitrides of the group III elements have led to commercially available, high-efficiency solid-state devices that emit green and blue light. Light-emitting diodes based on these materials should find applications in flat-panel displays, and blue and ultraviolet laser diodes promise high-density optical data storage and high-resolution printing.

The production of light has been of paramount importance to humankind since the dawn of civilization. From the use of fire to the incandescent electric lamp, the evolution of lighting technology has been characterized by the invention of increasingly brighter and more efficient light sources. The incandescent electric lamp was first demonstrated by Thomas Edison in 1879, and with the introduction of thin tungsten wire filaments at the turn of the century, light-bulb manufacture became a major industry. More powerful and efficient light sources followed, such as the neon lamps at the beginning of the century and the fluorescent tube in 1930. By the 1950s, the methods of lighting seemed well established, with filament and fluorescent lamps for interiors, neon lamps for exterior advertising and signs, and sodium discharge lamps for streets¹.

Electronic lighting technology made its appearance with the advent of semiconductors during the second part of the twentieth century. The achievement of high-efficiency electroluminescence from semiconductor devices in early 1962 led to light-emitting-diode (LED) technology². Later in the same year, light amplification by stimulated emission (laser action) in a semiconductor was demonstrated by four groups^{3,4}. Initially restricted to red light, and later extended to yellow, LEDs were first used in wrist watches and instrument panel indicators. The development of semiconductor diode lasers allowed the production of highly coherent bright light emitters that soon found applications in telecommunications (optical fibre networks), data storage (compact-disk technology), and document production (laser printers). LEDs operating in the red-to-yellow portion of the spectrum, with light emission efficiencies superior to those of incandescent lamps, became available in the early 1990s^{5,6}.

But it has proved difficult to extend these light sources into the short-wavelength region of the spectrum, from green to violet. Weak blue LEDs made from semiconducting SiC have been fabricated, but they are orders of magnitude less efficient than their red and yellow counterparts. The II–VI semiconducting compounds, especially those based on ZnSe, have shown some promise as blue and green light sources; but the facile formation of structural defects diminishes their lifetime and their potential usefulness for commercial applications. Much research has been done in the past three decades on the group III nitrides, comprising the Al–Ga–In–N alloys, which have the characteristic wide bandgap necessary for short-wavelength visible-light emission^{7–10}. In recent years research on nitride semiconductors has suddenly yielded dramatic advances, culminating in 1994 with the commercial introduction of extremely bright LEDs based on gallium nitride and operating in the green-to-ultraviolet range, with efficiencies superior to those of incandescent lamps and comparable to red and yellow LEDs. And these materials have now also been used in diode lasers operating at room temperature and emitting light in the blue-violet range, first

under pulsed conditions¹¹ in late 1995 and very recently under continuous operation¹².

With these technological advances, two particularly important areas of application are opened up. One is the ability to produce all three primary colours in LEDs, which implies applications in traffic lights, outdoor signs and large panel displays (see Box 1). The other is the availability of shorter-wavelength diode lasers, with the ability to project coherent radiation and to focus laser light into smaller spots (in the optical diffraction limit, the size of a focused spot is proportional to the wavelength of the light). This raises the possibility of high-density optical information storage and read-out. Green-to-ultraviolet lasers also bring closer to reality technologies such as full-colour photographic film printers, high-resolution laser printing, and projection television using full colour diode laser sources. Nitride semiconductors now look well placed to bring these possibilities to fruition.

The light-emitting diode

The light-emitting diode is the basic component for electronic lighting technology. It is a relatively simple semiconductor device that emits light when an electric current is passed through it. It consists of a back-to-back sandwich of p-type and n-type semiconductor materials (a p–n junction) characterized by a bandgap

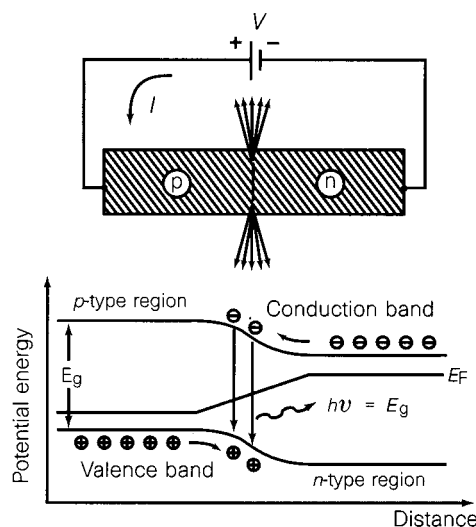


Figure 1 Schematic diagram of a light-emitting diode. The characteristics of the emitted light are related to the energy gap E_g between the conduction and valence band in the semiconductor. E_F is the Fermi energy.

Box 1 Achieving full-colour displays

The relevance and impact of the recent achievements in the III-V nitrides can be understood via the chromaticity diagram (see figure), which shows the relationship between the colours we see^{57,58}. The basis for the chromaticity diagram is the spectral sensitivity of the three types of cones in the human eye. Colorimetry therefore represents every colour by three parameters, corresponding to the output of each type of cone. As an example of how different colour mixtures can produce the same colour response, white light can be perceived in a number of ways. In one case, it may comprise a continuous spectrum of all visible wavelengths, as in the black-body emission from the Sun or an incandescent bulb. Alternatively, it may contain a series of more discrete spectral lines, for example as in the blue and yellow emissions from a fluorescent lamp. Although all these source colours appear white to the eye, their spectral compositions are substantially different.

The theoretical foundation of the chromaticity diagram is simply that any colour can be created by mixing three primary colours. Using a standardized set of primaries, the chromaticity diagram thus shows the amounts of each primary colour required to create another colour. The horizontal and vertical axes of the chromaticity diagram thus represent the fractions of two of the primaries. As the percentages of each of the three primary colours in the mixture must add to 100%, the fraction corresponding to the third primary is uniquely determined by specifying the other two. When the (x, y) coordinates of a point on the chromaticity are given, these correspond to the fractions of two primaries (with the remainder equivalent to the fraction of the third primary). Hence, the chromaticity diagram can show the colour of this particular combination, or any other combination, in a two-dimensional plot.

Any light source such as an LED, laser or lamp can be represented by a single point on the chromaticity diagram. Spectrally pure, monochromatic sources (such as a laser) are represented by points along the outer perimeter of the diagram. In this situation, when the peak wavelength (shown around the perimeter) and spectral width of such an emitter is specified, its coordinates on the chromaticity diagram are similarly determined. For instance, the coordinates of a pure blue emitter lie on the bottom left corner of the diagram (point 2). This corresponds to Nichia's super-bright blue InGaN LEDs. The output of pure red AlGaAs LEDs is identified by point 1.

When the spectral width of a source increases, and the emission is therefore less pure, the colour coordinates move towards the central (white) region of the diagram. This is indicated by point 3, which specifies Nichia's green 520 nm emission from an InGaN LED. Here, the spectral width is about 40 nm, or twice that of the blue LED. Consequently, the colour coordinate is displaced more from the perimeter of the chromaticity diagram. An extreme case of a spectrally broad emission spectrum is the black-body radiation associated with a body heated to about 10,000 K, which would appear white. This emission would be represented by a point in the central, white region of the chromaticity diagram, because this broadband emission is equivalent to a roughly equal mixture of all three primaries.

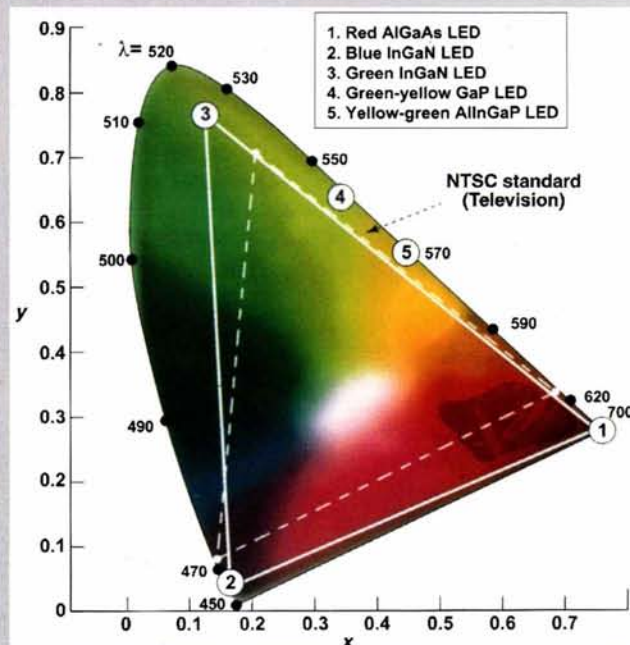
The chromaticity diagram also illustrates how two (or more) pure colours may be mixed to form an infinite palette of intermediate colours. Consider the case of a commercial red LED (point 1) and a green LED (point 3). When these two sources are combined, any colour on the line connecting them can be reproduced. For example, starting from the red colour specified by point 1, as a small amount of the green is mixed in, the colour will become orange. When the component red and green intensities are roughly equal (or midway between points 1 and 3), the colour of the mixed source is perceived as yellow. In a similar manner, yellow light may be combined with complementary blue light, to produce white. In fact, this combination of complementary yellow and blue colours is one approach to making solid-state white lamps; using a blue LED to excite phosphors is another alternative.

Similarly, when three sources are specified, they may be combined to produce all the colours contained within the triangular area defined by their coordinates on the chromaticity diagram. For example, the commercial red (point 1), green (point 3), and blue (point 2) LEDs may be combined to produce all the colours defined by the triangle shown. This demonstrates why the

recent development of bright blue and green emitters (taken with the pre-existing red LEDs) is so important. Using the pure red, green and blue emitters as primaries, nearly the entire chromaticity diagram is spanned. In other words, these three colours can be combined to generate nearly all colours, and so they form a minimum set of three sources which can be used to produce a true full-colour display. For comparison, the NTSC colour television standard based on luminescence from phosphor-coated cathode ray tubes is shown in the dashed triangle. From the figure it is observed that full-colour LEDs can provide a broader range of colours than those available on television screens.

Before 1994, only a narrow segment (yellow-green to red, or wavelength from 555 to 700 nm) of the colour diagram was accessible with high-efficiency LEDs. Although blue SiC LEDs were available before this, their efficiency was rather low, restricting their application. An optimum green wavelength of 520 nm was also not possible. The best green sources were in fact a yellow-green, fabricated from either nitrogen-doped GaP (point 4) or AlGaInP (point 3), a semiconductor alloy whose compositions can be adjusted to produce red, orange or yellow emission. Although AlGaInP is used to make the world's most efficient LEDs, which are yellow, it could not be extended to a shorter, green wavelength because it becomes an indirect bandgap semiconductor.

Today, most colours can not be produced by combining the red GaAlAs LEDs with the green and blue InGaN LEDs in a single full-colour LED lamp. In such a package, a sophisticated electronic controller adjusts the current to each LED, so that the combination of the red, green and blue outputs produces any desired colour. Similarly, when the nitride LEDs are evolved into green and blue semiconductor lasers, they could be combined with existing red lasers to create projection displays and colour film printers.



Box figure Chromaticity diagram from CIE (Commission Internationale de l'Éclairage) 1931. The monochromatic colours appear on the arch and their wavelength is indicated in nm. White light is located near the centre of the diagram. The colours of commercially available LEDs are shown. Note that the green and blue InGaN in combination with the red GaAlAs LEDs cover a large section of the chromaticity diagram, giving purer colour than the NTSC television standard (depicted by the dashed triangle). Colour range obtained from Photo Research Corporation, Chatsworth, California

E_g . This quantity determines the minimum energy required to excite an electron from the valence band to the conduction band (where it is mobile). Conversely, the bandgap also determines the energy of a photon produced when a conduction electron recombines with a hole in the valence band. When a current is passed through the diode, electrons in the conduction band flow across the junction from the n-doped side, and holes in the valence band flow from the p-doped side (Fig. 1). The result is that a significant number of electrons and holes recombine at the p–n junction, emitting light with an energy $h\nu \approx E_g$. A direct radiative transition involves only the emission of a photon, whereas an indirect transition requires a combination of a photon and a phonon (lattice vibration). Because of the required phonon interaction, the transition probability for a material with an indirect bandgap is usually much lower than in the case of a direct bandgap. The values of E_g for semiconductors emitting in the visible range of the spectrum are shown in Fig. 2 as a function of the interatomic bond length, which exerts a primary influence on the bandgap. The value of this lattice constant is critical for film fabrication technologies, as explained later.

All commercially available LEDs in the visible-to-infrared range (about 450–1,500 nm) are made of III–V compound semiconductor thin films, consisting of elements belonging to group III (aluminum, gallium and indium) and group V (nitrogen, phosphorus and arsenic) of the periodic table. Film deposition is achieved by vapour-phase techniques involving chemical reactions that produce thin films whose lattice structures are related to that of the substrate on which they are deposited (epitaxial growth). Red LEDs made from $\text{GaAs}_{1-x}\text{P}_x$ epitaxial layers were first introduced commercially by General Electric in 1962, following the work of Holonyak and Bevacqua³, and similar devices are still used today for low-intensity light applications. The evolution of LED performance is depicted in Fig. 3. Shorter-wavelength light emission requires a higher phosphorus content. However, for $x > 0.45$, $\text{GaAs}_{1-x}\text{P}_x$ has an indirect bandgap, causing the emission efficiency to drop

significantly. In the mid 1970s the introduction of nitrogen as a dopant enhanced the emission efficiency of $\text{GaAs}_{1-x}\text{P}_x$ LEDs by inducing a direct bandgap behaviour. Consequently, the performance of red LEDs was improved, and orange, yellow and green light emission became possible (the latter by nitrogen doping of GaP). But the light emission efficiencies of these LEDs were inferior to those of incandescent tungsten filament lamps, which, emitting at ~ 20 lumens per watt, have served as a point of reference for the performance of LEDs. (A lumen is the luminous flux intensity per solid angle of a point source with luminous intensity of one candela (cd).)

In the early 1980s, heterostructures involving junctions between materials with different bandgaps were introduced. In the late 1980s, double heterostructures (DHs), involving a smaller-bandgap semiconductor sandwiched by materials with larger bandgaps, increased the light-emission efficiency by improving electronic carrier confinement. As a result, red-emitting DH AlGaAs devices surpassed the performance of red-filtered incandescent lamps. Another order of magnitude improvement was achieved in the early 1990s with the introduction of red, orange and yellow LEDs based on AlInGaP (ref. 13). In the short-wavelength range, SiC blue LEDs also became available, but with much lower efficiencies; and work on II–VI compounds such as ZnSe was pursued intensively, aimed at developing brighter blue LEDs with sufficient durability for commercialization. The evolution of LEDs took a dramatic turn in late 1993, with the introduction by Shuji Nakamura of Nichia Chemical of highly efficient blue light emitters based on the group III nitride compounds. They were shortly followed by devices emitting in the green-to-violet range of the spectrum. This technology was made possible by persistent pioneering research over more than 20 years by Isamu Akasaki at Nagoya University (now at Meijo University). As a result, LEDs are available today covering all the primary colours of the spectrum, at competitive prices, and with energy efficiencies greater than those obtained from incandescent

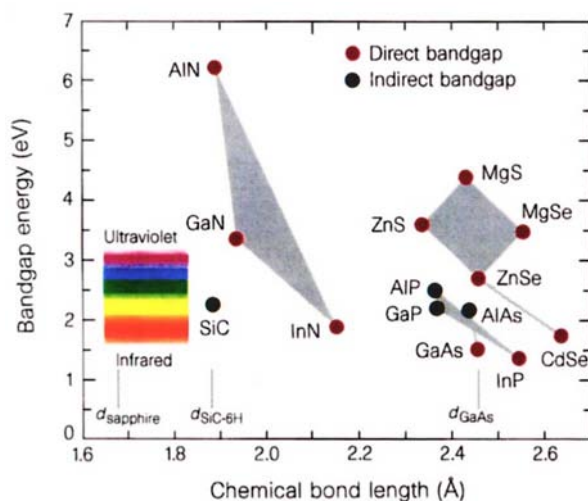


Figure 2 Bandgap and chemical bond lengths (d) of compound semiconductors that emit in the visible range of the electromagnetic spectrum. The visible spectrum is shown as related to the energy gap of the semiconductor. A direct-bandgap material involves only the emission of a photon per radiative event, whereas an indirect-bandgap material has radiative transitions involving a combination of a photon and a phonon (lattice vibration). The chemical bond length or interatomic separation is shown here, instead of the lattice parameter, to allow proper comparison of the hexagonal (wurtzite) nitrides with the cubic (zincblende) phosphides and arsenides. Also shown are the respective interatomic distances for relevant substrates (sapphire, SiC and GaAs).

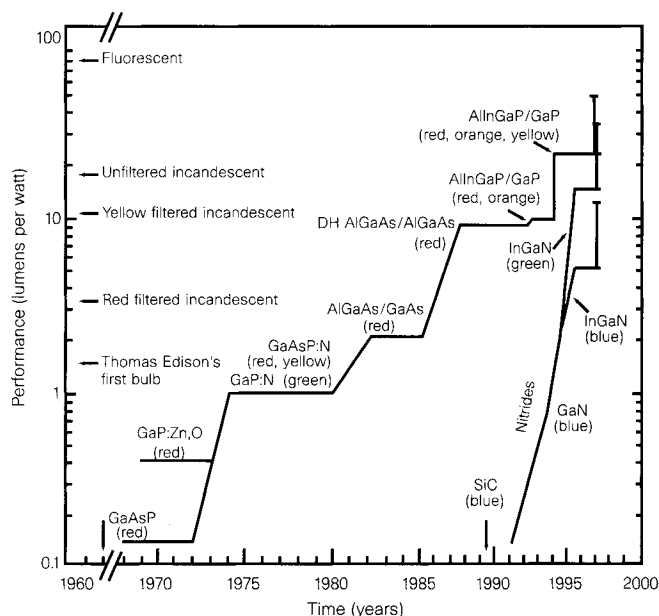


Figure 3 Evolution of visible LEDs. Over the past three decades, the arsenides and phosphides have produced increasingly brighter LEDs in the red-to-yellow range of the visible spectrum. In the past three years the nitrides have taken over the green-to-blue range. A lumen is the luminous flux intensity within unit solid angle of a point source with luminous intensity of one candela. An incandescent tungsten filament lamp emits about 20 lumens per watt. The bars for present day correspond to production (lower end) and laboratory (upper end) performance. (Updated version from report in ref. 6.)

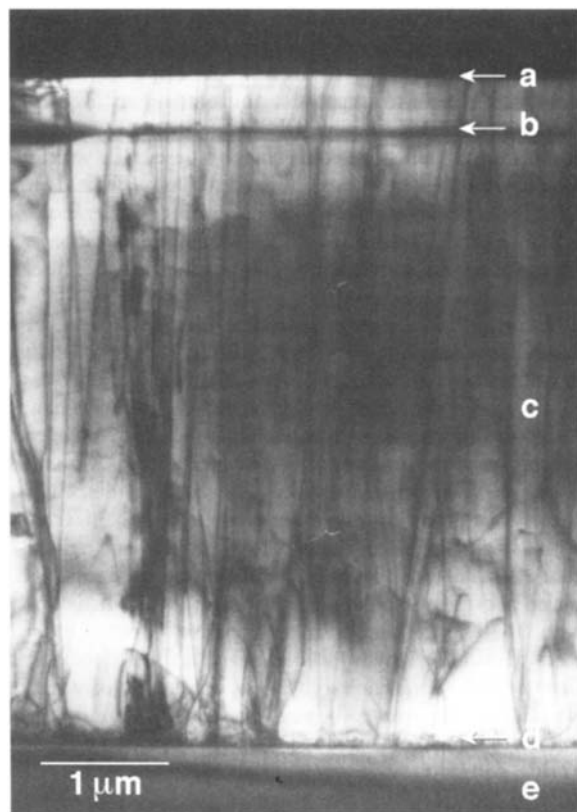


Figure 4 Cross-section transmission electron microscopy view of a high-efficiency Nichia blue LED. The dislocation density of 10^{10} cm^{-2} is remarkably high for such an efficient and durable device. Indicated in the figure are the top surface of the film (a), the active region (b), the main GaN film (c), the buffer layer (d), and the sapphire substrate (e).

tungsten-filament lamps.

The III–V nitrides

There are two fundamental reasons to choose nitrides for blue-light sources. Foremost is the large bandgap associated with the Al–Ga–In–N system, which encompasses the entire visible spectrum (Fig. 2). The $\text{Ga}_x\text{In}_{1-x}\text{N}$ ternary alloys, represented in the figure by the lower segment of the nitride triangle, have bandgaps ranging from 1.9 eV (red) for InN to 3.4 eV (ultraviolet) for GaN. The other main advantage of the nitrides over other high-bandgap semiconductors is the strong chemical bond which makes the material very stable and resistant to degradation under conditions of high electric currents and intense light illumination (as encountered in the active regions of the device). The strong chemical bonds account also for the difficulty in introducing dislocations once the material is grown, thus removing one of the main mechanisms of degradation observed in other III–V compounds.

Among the obvious disadvantages of the nitrides is the apparent absence of suitable substrates for film growth; the high growth temperatures made necessary by the strong chemical bonding in these materials limit the choice of substrates to those that are stable at such elevated temperatures. The two most commonly used are sapphire ($\alpha\text{-Al}_2\text{O}_3$) and silicon carbide (6H-SiC). The large differences in lattice parameter and in linear thermal expansion rates that occur both within the AlGaInN system itself and relative to these substrates meant that poor-quality epitaxial films were expected. The microstructure of films grown on sapphire, however, appears to have special properties that overcome these obstacles, as discussed later.

The III–V nitrides are not really new materials. AlN was reported in 1907 (ref. 14), and it is believed to be one of the first III–V compound semiconductors ever synthesized. The crystalline structure of GaN was described in 1937 (ref. 15) and growth of GaN crystalline films was reported in 1969 (ref. 16). Three important milestones were achieved in 1971. First, Pankove¹⁷ reported metal–

Table 1 Development of group III nitrides

Year	Event	Authors	Reference
1969	GaN by hydride vapour phase epitaxy	Maruska and Tietjen	16
1971	Metal-insulator-semiconductor LEDs	Pankove <i>et al.</i>	17
	GaN by MOCVD	Manasevit <i>et al.</i>	18
	Ultraviolet stimulated emission at 2 K	Dingle <i>et al.</i>	19
1974	GaN by sublimation	Matsumoto and Aoki	51
	GaN by MBE	Akasaki and Hayashi	52
1975	AlN by reactive evaporation	Yoshida <i>et al.</i>	53
1982	Synthesis (high pressure)	Karpinski <i>et al.</i>	54
1983	AlN intermediate layer (MBE)	Yoshida <i>et al.</i>	20
1986	Specular films using AlN buffer layers	Amano <i>et al.</i>	23
1989	p-type doping with Mg and LEEBI	Amano <i>et al.</i>	26
	GaN p–n junction LED	Amano and Akasaki	26
	InGaIn epitaxy (XRR = 100 arcmin)	Nagamoto <i>et al.</i>	55
1991	GaN buffer layer by MOCVD	Nakamura	37
1992	Mg activation by thermal annealing	Nakamura <i>et al.</i>	27
	High-brightness AlGaIn ultraviolet/blue LED (1.5%)	Akasaki and Amano	28
	InGaIn epitaxy (XRR = 5 arcmin)	Nakamura <i>et al.</i>	56
	High-brightness AlGaIn ultraviolet LEDs	Akasaki <i>et al.</i>	28
1993	InGaIn MQW structure	Nakamura <i>et al.</i>	29
	InGaIn/AlGaIn DH blue LEDs (1 cd)	Nakamura <i>et al.</i>	30
1994	InGaIn/AlGaIn DH blue-green LEDs (2 cd)	Nakamura <i>et al.</i>	31
1995	InGaIn QW blue, green and yellow LEDs	Nakamura <i>et al.</i>	33
	InGaIn SQW green LEDs (10 cd)	Nakamura <i>et al.</i>	34
	Blue laser diode, pulsed operation	Nakamura <i>et al.</i>	11
1996	Ultraviolet laser diode	Akasaki <i>et al.</i>	49
	Blue laser diode, pulsed operation	Itaya <i>et al.</i>	50
	Blue laser diode, c.w. operation	Nakamura <i>et al.</i>	12

Key steps in the development of the group III nitrides are listed with respective authors and references. The following abbreviations are used: low-energy electron beam irradiation (LEEBI), X-ray rocking curves (XRR), multiple quantum wells (MQW), quantum well (QW), single quantum well (SQW), double heterostructure (DH), continuous wave (c.w.).

insulator–semiconductor LEDs made of GaN thin films. Manasevit¹⁸ succeeded in growing GaN by the technique of metal-organic chemical vapour deposition (MOCVD) for the first time, using a method that is still used today. Finally, GaN needles were grown by Dingle¹⁹, who also reported stimulated light emission from this material at a temperature of 2 K, establishing the prospect of ultraviolet semiconductor injection lasers. The key developments during the following years are summarized in Table 1. The main obstacles to be overcome in making practical nitride-based devices involved difficulties in growing smooth films, in controlling n-type conductivity, in producing p-type materials and in growing AlGaIn and InGaIn alloys. Initial films were typically rough and fractured, features attributed to three-dimensional nucleation (a two-dimensional growth mode is required to produce smooth films) and the choice of substrates. These obstacles appeared insurmountable for many years, so the nitrides were largely ignored in favour of other large-bandgap materials.

Mastering epitaxy on foreign substrates is considered necessary to control the chemical purity of a semiconductor such as GaN, thus avoiding unwanted impurities which influence the optical and electronic character of the material. Similarly, during epitaxial growth it is desirable to minimize structural imperfections in the lattice such as dislocations and stacking faults, which are known to have adverse effects on optoelectronic properties of semiconducting materials (such as the quantum efficiency associated with light emission). In the search for improved properties, Yoshida²⁰ showed that an AlN intermediate layer between a sapphire substrate and an epitaxial GaN film increases the efficiency of cathodoluminescence (luminescence excited by a high-energy electron beam) of the films. Parallel developments, in particular in GaAs epitaxy on silicon^{21,22}, demonstrated that smooth epitaxial thin films on substrates with a large (> 4%) lattice mismatch could be achieved by the initial deposition of low-temperature nucleation (or buffer) layers. Although amorphous and therefore not of as high quality as the desired film, the buffer layers allowed for continuous coverage of the substrate and overcame the wetting obstacles related to surface and interface energetics. Using AlN low-temperature buffer layers on sapphire substrates, Akasaki and Amano²³ grew the first smooth surfaces of GaN films in 1986, demonstrating that two-dimensional nucleation in nitride epitaxy could be achieved and thereby bringing about a significant improvement in the electrical and luminescent properties of the films. Akasaki *et al.*²³ also demonstrated that silane (SiH₄) could be used for effective and controllable n-type doping.

Making p-type GaN films took a few more years. Magnesium was expected to be a good acceptor (p-type dopant); but although it was possible to incorporate significant concentrations of Mg in GaN, it did not effectively introduce hole charge carriers—the films were highly resistive. Similar observations in other materials such as GaAs and InP had established that unintentional hydrogen contamination played a strong role in passivating p-type dopants^{24,25}. In 1989, Amano and Akasaki demonstrated that low-energy electron-beam irradiation (LEEBI) activated magnesium-doped GaN films²⁶, converting them from a highly resistive state to one having a high ($\sim 10^{16}$ cm⁻³) concentration of holes. This p-type carrier density was sufficient to produce the first p–n junction blue LEDs²⁶. In 1992, Nakamura showed that superior hole concentrations could be achieved by thermal annealing after growth²⁷. These developments set the stage for brighter LEDs by Akasaki, Amano and others in 1992 (ref. 28). The development of extremely thin (quantum-well) layers of indium-rich nitrides by Nakamura brought increasingly superior performance^{29,30}, culminating with the announcement by Nichia Chemical Industries in November 1993 of the development of a bright blue LED and plans for mass production.

Blue and green LEDs

The first blue LEDs based on the III–V nitrides were made commercially available by Nichia in early 1994. They consisted of

an InGaIn/AlGaIn double heterostructure, fabricated by codoping Zn and Si into the InGaIn active layer²⁰. The output power was 3 mW at a forward current of 20 mA, with an external quantum efficiency (ratio of photons produced to the number of injected electrons) of 5.4% and a peak wavelength of 450 nm. Shortly thereafter, blue–green LEDs were fabricated by increasing the indium concentration in the InGaIn active layer, yielding a brightness of 2 cd (ref. 31).

Most surprising was that, compared with LEDs fabricated using other semiconductors, the light-emitting efficiencies of these GaN-based LEDs were remarkably insensitive to extended lattice defects such as dislocations in the films. Transmission electron microscopy (TEM) studies³² indicated that the material contained 10^{10} dislocations per square centimetre, more than six orders of magnitude above any other material used for LED fabrication. Figure 4 shows such a TEM cross-section image of a bright blue Nichia LED. Dislocations run in the growth direction (vertical), their density does not vary with distance from the sapphire interface, and they traverse the active region of the device. The fact that efficient GaN-based LEDs can be made despite such high densities of structural defects indicates a critical difference between the group III nitrides and other semiconductors.

Devices with longer peak emission wavelengths have been fabricated using extremely thin InGaIn layers (quantum wells) containing high concentrations of indium; these materials access the green and yellow range of the visible spectrum³³. In addition to producing longer wavelengths, quantum wells also narrow the spectral width of the emitted light, compared with the impurity-related transitions of the earlier DH LEDs, improving the monochromaticity. Bright green InGaIn single-quantum-well (SQW) LEDs have now been produced with an output power of 3 mW at a forward current of 20 mA, with a luminous intensity of 10 cd, an external quantum efficiency of 6.3%, and a peak wavelength of 520 nm (ref. 34). The green colour of the emitted light is much 'deeper' (more monochromatic) than that available from conventional GaP and AlInGaP, and the intensity is about 100 times higher.

Data on the reliability and lifetime of these devices are still limited. Long lifetimes have been reported, most in excess of 50,000 hours (lifetimes of $\sim 10^6$ hours > 100 years have been estimated S. Nakamura, personal communication), and there is no record of observed degradation mechanisms in the nitrides. Some reports even indicate that the material itself becomes more efficient with time, and that lamp failure is due to breakdown of the contacts or of the encapsulant but not of the semiconductor itself.

Growth and characteristics of the III–V nitrides

Several techniques, such as hydride vapour-phase epitaxy and molecular-beam epitaxy, have been successful in producing high-quality III–V nitride thin films. But as with other III–V compounds, MOCVD is the method of choice for commercial applications. The design of the growth reactor varies from vertical to horizontal flow configurations, and various gas sources have been used. The most commonly used gas-source precursors are trimethyl gallium (Ga(CH₃)₃) and ammonia (NH₃). Likewise, InGaIn and AlGaIn can be grown by the addition of trimethyl indium and trimethyl aluminum, respectively. Typical growth techniques today involve growth on [0001] sapphire substrates, with a low-temperature AlN or GaN buffer layer of thickness 10–40 nm deposited at about 550 °C. This is followed by GaN epitaxy at 700–1,100 °C. The initial buffer layer is amorphous and undergoes solid-phase crystallization during the heating step prior to epitaxy. TEM images of the substrate/buffer layer region (Fig. 5a) indicate an abrupt crystalline transition between the two lattices³⁵. The lack of coherence (continuity) between the sapphire and the nitride is also apparent; high-quality epitaxy is evidently achieved without such one-to-one correspondence between the crystalline lattices. Slight variations in the orientation of the epitaxial GaN layer are accommodated by

dislocations. This results in a columnar structure consisting of single crystal domains with diameters of under one micrometre, and with a distribution in lattice orientation (columnar tilt or twist) of about 5 arcmin. Spatially inhomogeneous cathodoluminescence correlates well with the columnar structure and the distribution of dislocations in the material, indicating a connection between microstructure and light emission³⁶. This microstructure results in high-quality GaN films exhibiting smooth surface morphologies and electron mobilities of about $600 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ at room temperature^{23,37}. The carrier concentration of n-type GaN films can be controlled between $1 \times 10^{17} \text{ cm}^{-3}$ and $2 \times 10^{19} \text{ cm}^{-3}$ using silicon as dopant³⁷. Similarly, using magnesium and activating the doping process by thermal annealing²⁷, p-type GaN can be produced with hole concentrations of the order of $1 \times 10^{18} \text{ cm}^{-3}$.

In addition to sapphire, silicon carbide is also used for the growth of the nitrides. The lattice structure, interatomic separation and thermal expansion rates of the nitrides are much more similar to those of SiC than to those of sapphire^{38,39}. These close similarities produce highly coherent changes in the atomic arrangement between SiC and the nitrides, as seen in lattice images of the interface between these two materials (Fig. 5b). In view of these similarities, it is surprising that so far devices fabricated from nitride epitaxy on SiC have inferior properties to those deposited on sapphire. One possible explanation may be the difficulty of preparing atomically flat SiC surfaces. Another is that dislocations do not compromise the performance of the materials; indeed, they even appear to be beneficial, by providing local-strain relaxation mechanisms which allow high optoelectronic performance. Epitaxy on bulk GaN single crystals has been demonstrated⁴⁰, with dislocation densities below 10^6 cm^{-2} . The poor availability of sufficiently large GaN single crystals, however, limits the practicality of this choice of substrate.

Diode lasers

The main focus of research on III–V nitrides is now the realization of a current-injected laser diode that can be operated under continuous-wave (c.w.) conditions at room temperature. Optically pumped stimulated emission has been reported in GaN (refs 41, 42), InGaN (refs 43, 44), AlGaIn/InGaIn double heterostructures⁴⁵, and GaN/AlGaIn double heterostructures⁴⁶. But despite the

established success of nitride LEDs by 1995, there was still much uncertainty concerning the feasibility of injection lasers. In particular, it was not clear that material with such a heavily faulted crystal structure could survive the extremely high current density required for generating optical gain in a laser, roughly two orders of magnitude greater than LED current. In late 1995 and early 1996, however, Nakamura demonstrated violet–blue injection laser action in InGaIn/GaN/AlGaIn-based heterostructures under pulsed conditions on {0001} and {11 $\bar{2}$ 0} sapphire substrates^{11,47} (Fig. 6).

A laser diode differs from an LED in that it must provide optical gain, guide the emitted photons and also incorporate a resonant cavity for feedback. The layer stack that makes up a laser diode structure (Fig. 7) is thus more sophisticated than that of an LED, as it must accomplish these several additional functions. As in an LED (Fig. 1), the heterostructure must energetically confine injected carriers into an active region. But because lasers operate at much higher current densities than LEDs, greater confinement is required. The laser heterostructure also forms a slab waveguide, designed to concentrate the optical mode over the p–n junction (active region) in order to maximize the optical gain.

In a semiconductor, optical gain is realized when a high population of electrons and holes is injected into the active region. The lasing threshold is reached when the optical gain is sufficient to overcome all the optical losses. Pumping mechanisms for semiconductor lasers include optical- and electron-beam excitation, or, in the case of a laser diode, carrier injection across a p–n junction. Accordingly, the active region of a laser diode is located precisely at the p–n junction, so that when the diode is forward biased, the charge carriers are injected into the active region. Because the active region has a smaller bandgap than the surrounding layers, the electrons and holes are energetically confined to this region; and because of its small dimensions, a high injected carrier density is more easily achieved.

The active layer of high-performance devices are typically ultra-thin (20–50 Å) layers. Such layers are called quantum wells, because their width is less than the de Broglie wavelength of a conduction electron. This space quantization in one dimension gives rise to several beneficial effects. First, the available electron states become quantized, leading to an overall reduction in the density of states,

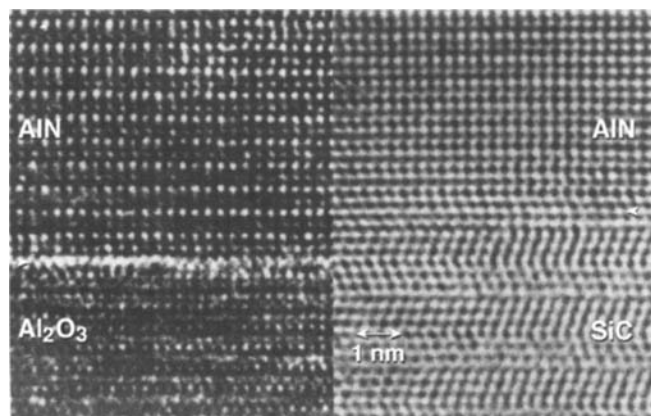


Figure 5 Atomic arrangement at the substrate/nitride interface is critical for high-quality material. Left, nitrides grown on sapphire are epitaxial in spite of the large difference in lattice parameter and crystalline structure. Right, nitride epitaxy on SiC is characterized by coherent interfaces due to the similarity between the lattice structures. In both cases the interfaces are atomically flat and abrupt, indicative of the stability and ease of growth of these materials.



Figure 6 Laser diode operating under pulsed current injection of 1.4 A at room temperature⁴⁷ (courtesy of S. Nakamura, Nichia).

thereby requiring that fewer carriers be injected to reach optical gain. Second, the quantization influences the electron and hole wavefunctions, thereby making the gain anisotropic with respect to polarization. This anisotropy favours one polarization over the other, with a consequent increase in optical gain. For these reasons, a quantum-well (or multiple-quantum-well) active region is superior to a thicker active region, which behaves more like a bulk material.

In a semiconductor laser, a resonant cavity is typically realized by cleaving end-facets on a laser diode and relying on reflection from these facets. Because the entire wafer and epitaxial layers form a single crystal, the facets are thus guaranteed to be perfectly parallel and vertical. Unfortunately, the cleavage planes of GaN and sapphire are not necessarily aligned, making cleaved facets difficult. Consequently, the formation of a resonant cavity is difficult for nitrides. Whereas the first nitride laser had very rough, plasma-etched facets, recently they have been formed by more traditional cleaving⁴⁷. In this case, cleaved facets were possible because the layers were grown on {1120} sapphire rather than {0001} substrates: this makes the cleavage planes of the GaN and sapphire nearly aligned and perpendicular to the growth plane.

The threshold current density of nitride lasers is still very high; the best reported values⁴⁸ are about $3,000 \text{ A cm}^{-2}$, more than an order of magnitude higher than those of red and infrared laser diodes. This, coupled with their relatively high electrical resistance, makes heat generation a problem. Lower thresholds are likely, however, as the layer structures and resonators are optimized. For example, the first nitride injection laser had an active region containing 26 InGaN quantum wells, each being only 25 Å thick and separated by 75-Å barriers¹¹. Although the requirement for multiple quantum wells might be aimed at achieving sufficient optical gain, this was still an unusually high number of quantum wells. Nevertheless, since this first demonstration, Nakamura's structures have evolved towards fewer (3–5) quantum wells and better-quality facets⁴⁸. Likewise, Akasaki has produced a single-quantum-well laser that operates in the ultraviolet⁴⁹. Also notable is the demonstration of a pulsed nitride laser by Itaya *et al.* at Toshiba⁵⁰. The most impressive achievement so far, in late 1996, is Nakamura's operation of a nitride laser continuously at room temperature for 35 hours, under a constant output power of 1.5 mW per facet, with operating currents in the range of 100–280 mA (ref. 12).

InGaN segregation may play a critical role in the mechanism by

which optical gain is produced in the nitrides. When the InGaN alloy becomes unstable and separates into regions of different compositions, the regions of high-indium content tend to confine injected carriers, as a consequence of their lower bandgap energy. Thus, the InGaN segregation may lead to the formation of quantum dots, providing space quantization in three dimensions. Furthermore, the optical gain may also be excitonic (involving bound electron–hole pairs, rather than free charge carriers) in the nitrides. Indeed, unlike in red or infrared lasers where excitons cannot survive at room temperature and under the high injected carrier densities associated with lasing, the optical gain in the nitrides may arise from exciton recombination; or at least the gain is somehow 'Coulomb-enhanced', where the attractive interaction between the electron and hole plasmas becomes important. If the gain is excitonic, even slight InGaN composition variations could be significant, as they would induce weak potential fluctuations in the quantum well plane, leading to exciton localization. These issues remain to be explored fully.

Applications

The high efficiency and longevity of the III–V nitrides satisfy the requirements for short-wavelength visible LEDs. Red traffic signals are currently produced by filtering the red portion of incandescent lamps, wasting about 85% of the emitted light. To obtain green and amber, nearly 30% of the light must be filtered out. Power efficiency is an important factor for traffic lights, as they typically operate continuously throughout the day. In addition, LEDs have lifetimes of over 5 years whereas incandescent lamps need replacement a few times per year. The combined savings in consumed energy and in replacement costs make the use of LEDs in traffic signals economically desirable.

Large full-colour outdoor electroluminescent displays using green and blue InGaN LEDs already exist in various cities, and more are expected as the relatively large cost is compensated by their impact. Automobile interior panel lighting is also an important market for full-colour LEDs. As manufacturing costs drop with increased mass production, similar considerations will apply to all types of lighting: home and office lighting might also be generated by LEDs, installed as permanent fixtures to last the lifetime of the building.

Short-wavelength injection lasers are highly desirable for applications such as ultra-high-density optical data storage, high-resolution printing, and for defence-related underwater communications

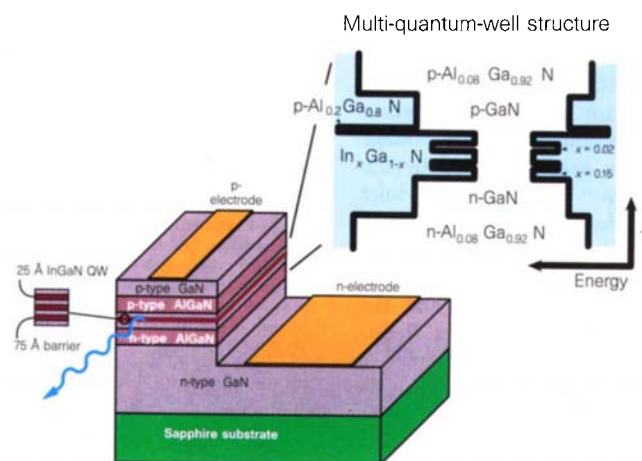


Figure 7 Diagram of the structure of InGaN multiple quantum well (MQW) diode laser⁴⁷. Inset shows the energy band diagram associated with the MQW structure. Carrier confinement is achieved by the adjacent layers of wider bandgap GaN

and the AlGaIn cladding layers. In addition, the p-type $\text{Al}_{0.2}\text{Ga}_{0.8}\text{N}$ layer immediately above the MQW may play a critical role in carrier confinement.

requiring green lasers. The first of these, optical data storage, is a dominant motivation for blue-laser development. This can be understood by noting that in the diffraction limit, the minimum spot size for a given optical system is proportional to wavelength λ ; thus the optical recording areal densities scale as $1/\lambda^2$. Furthermore, as the data are more tightly stored on such a disk, the data retrieval rate is also increased. Current optical disks are based on infrared light ($\lambda = 780$ nm) and are in the process of upgrading to red light ($\lambda = 650$ nm). With the recent demonstrations of blue laser diodes ($\lambda = 450$ nm), it seems that the commercial prospects are promising, particularly for digital video disk technology. The figure of merit of data-storage technologies is the areal density; currently, commercial devices offer 500 Mbit in^{-2} . With the advent of blue lasers, this figure will be increased by a factor of four as a result of the shorter wavelength. Similar considerations also apply to high-resolution printers, where a small focused spot translates into greater resolution. Other applications on the horizon include full-colour film printers, projection television and biomedical uses.

Challenges

Perhaps the main challenge in the field of the nitrides is to increase the lifetime of diode lasers operating at room temperature. The short lifetimes are caused by heating during operation, a consequence of the high threshold current along with the relatively high electrical resistance of nitride lasers. The heating problem may therefore be ameliorated either by reducing the threshold current or by lessening the diode's electrical resistance. Although theories indicate that the threshold currents are relatively high for the nitrides (higher than for red or infrared lasers), it is difficult to assess how this depends on the highly dislocated crystal structure. For example, defects may be responsible for other recombination pathways, or they may scatter light out of the optical cavity. Better lattice-matched substrates might lower thresholds by eliminating crystal defects; but on the other hand, the fact that nitride LEDs already show high quantum efficiencies despite their high dislocation densities may indicate that dislocations are not seriously detrimental (although scattering is not a problem in LEDs, it is a serious issue in lasers). In fact, dislocations can under certain conditions actually be favourable, as they may provide mechanisms by which thermal strain is relieved.

Reducing the electrical resistance of nitride laser diodes would similarly bring closer the goal of reliable, continuous, room-temperature operation. At present, the nitride diodes are relatively resistive, leading to severe heating at the high current densities required of lasing. Of the many component layers making up the diode, the metal contact to the p-type layers is especially resistive, usually dominating the total series resistance of the diode. This is so because of the limited p-type doping possible in the nitrides, and the high Schottky energy barrier associated with the interface between a metal and a high-bandgap semiconductor. Energy conservation requires that in order to support injection, the diode voltage should be at least equal to the photon energy divided by the electronic charge, corresponding to about 3 V for a 420 nm violet photon. In red and infrared laser diodes, the diode voltage at threshold is typically a few tenths of a volt above this. In contrast, the diode voltage of the first nitride laser was about 35 V. This included the drop of ~ 3 V across the p-n junction, and the remaining ~ 30 V across the metal-semiconductor contact. Reducing the p-contact resistance, either by higher doping or alternative contact materials, is therefore critical to overcoming the heating problem sufficiently to obtain long-lived continuous operation of nitride lasers at room temperature.

Spanning a broader range of wavelengths is also an aim of nitride laser development, driven by numerous potential applications. For example, communication through sea water is performed optimally at green wavelengths (~ 505 nm), whereas optical data storage demands as short a wavelength as possible while remaining con-

sistent with the spectral range over which inexpensive optical systems maintain their transparency.

Expanding nitride laser operation from the ultraviolet or violet to longer wavelengths is made challenging primarily by the growth of high-indium-content InGa_N alloys for the active region. In principle, the composition of an InGa_N alloy active region can be adjusted to give emission throughout the entire visible spectrum. For instance, a thin (20–50 Å) In_{0.3}Ga_{0.7}N layer provides blue emission, whereas increasing the indium content from 30 to 50% shifts the wavelength to green. However, because the lattice parameter of InGa_N alloys is larger than for Ga_N, only thin layers can be grown before structural defects such as dislocations or alloy segregation (where the strain within a uniform InGa_N alloy film causes it to dissociate into regions of highly disparate compositions) arise. These limitations of InGa_N growth are still being investigated.

With the advent of the nitride-based semiconductors, the long search for bright solid-state sources covering the full range of the visible spectrum seems to be over. Green and blue nitride light-emitting diodes are now available commercially with surprisingly high efficiencies and long lifetimes. Blue and ultraviolet diode lasers have been demonstrated under pulsed conditions, and continuous operation of nitride lasers for tens of hours is already possible. These exciting recent developments have been made possible only by decades of persistent research, and by the unique characteristics of the nitride-based semiconductors. They promise a bright future for photonic technology. □

F. A. Ponce and D. P. Bour are at Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California 94304, USA.

1. Lighting and lighting devices in *The New Encyclopedia Britannica* 15th Edn, Vol. 23, 29–38 (Encyclopaedia Britannica Inc., Chicago, 1986).
2. Black, J., Lowcockwood, H. & Mayburg, S. Recombination radiation in GaAs. *J. Appl. Phys.* **34**, 178–180 (1963).
3. Holonyak, H. Jr & Bevacqua, S. F. Coherent (visible) light emission from Ga(As)_{1-x}(P)_x junctions. *Appl. Phys. Lett.* **1**, 82–83 (1962).
4. Dupuis, R. D. An introduction to the development of the semiconductor laser. *IEEE J. Quant. Electr.* **QE-23**, 651–657 (1987).
5. Craford, M. G. LEDs challenge the incandescents. *IEEE Circuit Devices* **8**, 25–29 (1992).
6. Craford, M. G. & Steranka, F. M. Light Emitting Diodes, in *Encyclopedia of Applied Physics* Vol. 8, 485–514 (VCH, New York, 1994).
7. Akasaki, I. & Amano, H. Wide-gap column-III nitride semiconductors for UV/blue light emitting devices. *J. Electrochem. Soc.* **141**, 2266–2271 (1994).
8. Morkoc, H. et al. Large-band-gap SiC, III–V nitride, and II–VI ZnSe-based semiconductor device technologies. *J. Appl. Phys.* **76**, 1363–1398 (1994).
9. Mohammad, S. N., Salvador, A. A. & Morkoc, H. Emerging gallium nitride based devices. *Proc. IEEE* **83**, 1306–1355 (1995).
10. Neumayer, D. A. & Ekerdt, J. G. Growth of group III nitrides: A review of precursors and techniques. *Chem. Mater.* **8**, 9–25 (1996).
11. Nakamura, S. et al. InGa_N-based multi-quantum-well structure laser diodes. *Jpn. J. Appl. Phys.* **35**, L74–L76 (1996).
12. Nakamura, S. Characteristics of InGa_N multi-quantum-well-structure laser diodes. *Mater. Res. Soc. Proc.* **449**, 1135–1142 (1996).
13. Kuo, C. P., Fletcher, R. M., Osentowski, T. D., Lardizabal, M. C. & Craford, M. G. High performance AlGaInP visible light-emitting diodes. *Appl. Phys. Lett.* **57**, 2937–2939 (1990).
14. Fichter, F. Über Aluminiumnitrid. *Z. Anorg. Chem.* **54**, 322–327 (1907).
15. Tirmann, J. V. & Schdanov, H. S. The crystalline structure of Ga_N. *Acta Physicochim. URSS* **6**, 306 (1937).
16. Maruska, H. P. & Tietjen, J. J. The preparation and properties of vapor-deposited single-crystalline Ga_N. *Appl. Phys. Lett.* **15**, 327–329 (1969).
17. Pankove, J. I., Miller, E. A. & Berkeyheiser, J. E. Ga_N electroluminescent diodes. *RCA Rev.* **32**, 383–392 (1971).
18. Manasevit, H. M., Erdmann, F. M. & Simpson, W. I. The use of metalorganics in the preparation of semiconductor materials. *J. Electrochem. Soc.* **118**, 1864–1868 (1971).
19. Dingle, R., Shalkev, K. I., Leheny, R. F. & Zetterstrom, R. B. Stimulated emission and laser action in gallium nitride. *Appl. Phys. Lett.* **19**, 5–7 (1971).
20. Yoshida, S., Misawa, S. & Gonda, S. Improvements on the electrical and luminescent properties of reactive molecular beam epitaxially grown Ga_N films by using AlN-coated sapphire substrates. *Appl. Phys. Lett.* **42**, 427–429 (1983).
21. Chang, C.-A., Takaoka, H., Chang, L. I. & Esaki, L. Molecular beam epitaxy of AlSb. *Appl. Phys. Lett.* **40**, 983–985 (1982).
22. Wang, W. I. Molecular beam epitaxial growth and material properties of GaAs and AlGaAs on Si(100). *Appl. Phys. Lett.* **44**, 1149–1151 (1984).
23. Amano, H., Sawaki, N., Akasaki, I. & Toyoda, Y. Metalorganic vapor phase epitaxial growth of a high quality Ga_N film using an AlN buffer layer. *Appl. Phys. Lett.* **48**, 353–355 (1986).
24. Johnson, N. M., Burnham, R. D., Street, R. A. & Thornton, R. L. Hydrogen passivation of shallow-acceptor impurities in p-type GaAs. *Phys. Rev. B* **33**, 1102–1105 (1986).
25. Antell, G. R. et al. Passivation of zinc acceptors in InP by atomic hydrogen coming from arsine during metalorganic vapor phase epitaxy. *Appl. Phys. Lett.* **53**, 758–760 (1988).
26. Amano, H., Kito, M., Hiramatsu, K. & Akasaki, I. P-type conduction in Mg-doped Ga_N treated with low-energy electron beam irradiation (LEEBI). *Jpn. J. Appl. Phys.* **28**, L2112–L2114 (1989).
27. Nakamura, S., Mukai, T., Senoh, M. & Iwasa, N. Thermal annealing effects on p-type Mg-doped Ga_N films. *Jpn. J. Appl. Phys.* **31**, L139–L142 (1992).

28. Akasaki, I., Amano, H., Itoh, K., Koide, N. & Manabe, K. GaN-based ultraviolet/blue light emitting devices. *Inst. Phys. Conf. Ser.* **129**, 851–856 (1992).
29. Nakamura, S., Mukai, T., Senoh, M., Nagahama, S. & Iwasa, N. $\text{In}_x\text{Ga}_{1-x}\text{N}/\text{In}_y\text{Ga}_{1-y}\text{N}$ superlattices grown on GaN films. *J. Appl. Phys.* **74**, 3911–3915 (1993).
30. Nakamura, S., Mukai, T. & Senoh, M. Candela-class high-brightness InGaN/AlGaIn double-heterostructure blue-light-emitting diodes. *Appl. Phys. Lett.* **64**, 1687–1689 (1994).
31. Nakamura, S., Mukai, T. & Senoh, M. High-brightness InGaN/AlGaIn double heterostructure blue-green light-emitting diodes. *J. Appl. Phys.* **76**, 8189–8191 (1994).
32. Lester, S. D., Ponce, F. A., Craford, M. G. & Steigerwald, D. A. High dislocation densities in high efficiency GaN-based light-emitting diodes. *Appl. Phys. Lett.* **66**, 1249–1251 (1995).
33. Nakamura, S., Senoh, M., Iwasa, N. & Nagahama, S. High-brightness InGaIn blue, green and yellow light-emitting diodes with quantum well structures. *Jpn. J. Appl. Phys.* **34**, L797–L799 (1995).
34. Nakamura, S., Senoh, M., Iwasa, N., Nagahama, S., Yamada T. & Mukai, T. Superbright green InGaIn SQW structure LEDs. *Jpn. J. Appl. Phys.* **34**, L1332–L1335 (1995).
35. Ponce, F. A., Major, J. S. Jr, Plano, W. E. & Welch, D. F. Crystalline structure of AlGaIn epitaxy on sapphire using AlN buffer layers. *Appl. Phys. Lett.* **65**, 2302–2304 (1994).
36. Ponce, F. A., Bour, D. P., Götz, W. & Wright, P. I. Spatial distribution of the luminescence in GaN thin films. *Appl. Phys. Lett.* **68**, 57–59 (1996).
37. Nakamura, S. GaN growth using GaN buffer layer. *Jpn. J. Appl. Phys.* **30**, L1705–L1707 (1991).
38. Ponce, F. A., Krusor, B. S., Major, J. S. Jr, Plano, W. E. & Welch, D. F. Microstructure of GaN epitaxy on SiC using AlN buffer layers. *Appl. Phys. Lett.* **67**, 410–412 (1995).
39. Ponce, F. A., Van de Walle, C. G. & Northrup, J. E. Atomic arrangement at the AlN/SiC interface. *Phys. Rev. B* **53**, 7473–7478 (1996).
40. Ponce, F. A. *et al.* Homoepitaxy of GaN on polished bulk single crystals by metalorganic chemical vapor deposition. *Appl. Phys. Lett.* **68**, 917–919 (1996).
41. Amano, H., Asahi, T. & Akasaki, I. Stimulated emission near ultraviolet at room temperature from a GaN film grown on sapphire by MOVPE using an AlN buffer layer. *Jpn. J. Appl. Phys.* **29**, L205–L206 (1990).
42. Zubrilov, A. S. *et al.* Spontaneous and stimulated emission from photopumped GaN grown on SiC. *Appl. Phys. Lett.* **67**, 533–535 (1995).
43. Khan, M. A., Krishnakutty, S., Skogman, R. A., Kuznia, J. N. & Olson, D. T. Vertical-cavity stimulated emission from photopumped InGaIn/GaN heterojunctions at room temperature. *Appl. Phys. Lett.* **65**, 520–521 (1994).
44. Kim, S. T., Amano, H. & Akasaki, I. Surface-mode stimulated emission from optically pumped GaInN at room temperature. *Appl. Phys. Lett.* **67**, 267–269 (1995).
45. Amano, H. *et al.* Room-temperature violet stimulated emission from optically pumped AlGaIn/GaInN double heterostructure. *Appl. Phys. Lett.* **64**, 1377–1379 (1994).
46. Aggarwal, R. L., Maki, P. A., Molnar, R. J., Liao, Z.-L. & McInnis, I. Optically pumped GaN/AlGaIn double heterostructure ultraviolet laser. *J. Appl. Phys.* **79**, 2148–2150 (1996).
47. Nakamura, S. *et al.* InGaIn multi-quantum-well-structure laser diodes with cleaved mirror cavity facets. *Jpn. J. Appl. Phys.* **35**, L217–L219 (1996).
48. Nakamura, S. *et al.* Ridge-geometry InGaIn multi-quantum-well-structure laser diodes. *Appl. Phys. Lett.* **69**, 1477–1479 (1996).
49. Akasaki, I. *et al.* Shortest wavelength semiconductor laser diode. *Electron. Lett.* **32**, 1105–1106 (1996).
50. Itaya, K. *et al.* Room temperature pulsed operation of nitride based multi-quantum-well laser diodes with cleaved facets on conventional *c*-face sapphire substrates. *Jpn. J. Appl. Phys.* **35**, L1315–L1317 (1996).
51. Matsumoto T. & Aoki, M. Temperature dependence of photoluminescence from GaN. *Jpn. J. Appl. Phys.* **13**, 1804–1807 (1974).
52. Akasaki, I. & Hayashi, I. Research on blue emitting devices. *Ind. Sci. Technol.* **17**, 48–52 (1976) (in Japanese).
53. Yoshida, S., Misawa, S. & Itoh, A. Epitaxial growth of aluminum nitride films on sapphire by reactive evaporation. *Appl. Phys. Lett.* **26**, 461–462 (1975).
54. Karpinski, J., Porowski, S. & Miotkowska, S. High pressure vapor growth of GaN. *J. Cryst. Growth* **56**, 77–82 (1982).
55. Nagamoto, T., Kuboyama, T., Minamino, H. & Omoto, O. Properties of $\text{Ga}_{1-x}\text{In}_x\text{N}$ films prepared by MOVPE. *Jpn. J. Appl. Phys.* **28**, L1334–L1336 (1989).
56. Nakamura, S. & Mukai, T. High-quality InGaIn films grown on GaN films. *Jpn. J. Appl. Phys.* **31**, L1457–L1459 (1992).
57. Hunt, R. W. G. *Measuring Colour* 1–313 (Ellis Horwood, London, 1995).
58. Wicks, G. W. The colour of light. *Compound Semiconductor* **1**, 39–40 (1995).

Acknowledgements. We acknowledge partial support from the Department of Commerce Advanced Technology Program, and from ARPA.

Correspondence should be addressed to F.A.P. (e-mail: ponce@parc.xerox.com).

Radical fringe positions the apical ectodermal ridge at the dorsoventral boundary of the vertebrate limb

Concepción Rodríguez-Esteban*, John W. R. Schwabe*, Jennifer De La Peña, Bryon Foy, Brian Eshelman & Juan Carlos Izpisua Belmonte

Gene Expression Laboratory, The Salk Institute, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

* Joint first authors.

Vertebrate limb outgrowth requires a structure called the apical ectodermal ridge, formation of which follows the previous establishment of the dorsoventral limb axis. *Radical fringe* is expressed in the dorsal ectoderm before the ridge appears, and is repressed by *Engrailed-1*, which is expressed in the ventral ectoderm. Misexpression of these genes indicates that a ridge is formed wherever there is a boundary between cells expressing and not expressing *Radical fringe*. Thus, as in *Drosophila*, *Radical fringe* positions the ridge at the dorsoventral limb boundary.

Regulation of vertebrate development depends on communication between different cells and tissues. Vertebrate limbs arise as a consequence of interactions between mesodermal and ectodermal cells established very early during embryogenesis. These interactions lead to the appearance, some days later, of a limb bud composed of a ball of undifferentiated mesenchymal cells surrounded by an ectodermal jacket. Transplantation experiments have identified three major signalling centres that establish the three major axes of the developing limb: the polarizing region, a group of mesenchymal cells located at the posterior of the limb, mediates patterning along the anteroposterior axis^{1,2}; the ectoderm covering the limb 'instructs' the underlying mesoderm and thereby mediates dorsoventral patterning^{3,4}; and the apical ectodermal ridge (AER), a specialized epithelium that arises at the junction of the future dorsal and ventral ectoderm^{5,6}, integrates some of the information provided by the two other centres and in turn transmits new signals that result in limb outgrowth.

Although it is becoming clear how the anteroposterior axis is established and elaborated⁷⁻¹⁰, only recently have the mechanisms through which the dorsoventral and proximodistal axes are set up begun to be unravelled. Genes known to specify dorsal cell fate include *Wnt-7a* and *Lmx-1*, which are restricted to the dorsal limb ectoderm and mesoderm, respectively¹¹⁻¹⁴. *En-1*, which is expressed in the ventral ectoderm, has been shown to play a role in establishing ventral cell fate¹⁵. Members of the fibroblast growth factor family, *Fgf-2*, *Fgf-4* and *Fgf-8*, play a key role during limb development as they not only mediate interactions between the mesoderm and ectoderm that allow outgrowth of the limb, but they can also induce formation of the AER and maintain limb outgrowth when the AER is removed¹⁶⁻²⁴.

Here we present the cloning of a chick gene, *Radical fringe* (*R-fng*), that is homologous to the *Drosophila fng* gene²⁵. We show that *R-fng* plays an important role in vertebrate limb outgrowth; in particular, it directs the formation and positioning, at the dorsoventral limb boundary, of one of the key organizer centres of vertebrate limb development, the AER. Further, we have investigated whether *R-fng* expression regulates or is regulated by the genes *Wnt-7a* and *En-1* that establish this boundary. Our results suggest that correct *R-fng* expression requires, but does not influence, the prior establishment of the dorsoventral boundary.

It has been shown previously in *Drosophila* that *fng* is required in an analogous fashion for the formation of the wing margin and hence the wing itself²⁵. The existence of evolutionary parallels

between the vertebrate AER and the *Drosophila* wing margin is supported by our finding that genes which function downstream of *fng* in *Drosophila* (*Serrate* and *Notch*²⁶⁻³⁰), are also conserved and expressed in the vertebrate AER. Taken together these results suggest that some of the mechanisms leading to limb formation in insects and vertebrates have been preserved. Nevertheless, our data also indicate that not all the molecules involved have been conserved, and this raises questions as to the causes and consequences of changing a subset of players within a specific developmental program.

R-fng* is homologous to *Drosophila fng

Chick genes homologous to the *Drosophila fng* gene were cloned by screening a limb bud cDNA library with a probe complementary to the *Drosophila* gene. One of these genes shows a striking expression pattern in the limb. This clone has particularly high homology to the *Xenopus radical fringe* gene³¹, and hence we call it chick *Radical fringe* (*R-fng*; see Supplementary Information for amino-acid sequence alignment). *R-fng* encodes a protein of 372 amino acids, with a putative amino-terminal signal sequence and a tetrabasic cleavage site, suggesting that it is a secreted signalling molecule like the related *Xenopus* and *Drosophila* proteins. Weak but significant similarity³² of *fng* proteins to the *Lex1* family of biosynthetic enzymes³³ suggests that *R-fng* may function as a glycosyltransferase (see alignment in Supplementary Information). This homology indicates that *R-fng* may operate through the recognition, and perhaps modification, of carbohydrate moieties linked to either extracellular proteins or lipids.

***R-fng* is expressed in dorsal ectoderm and AER**

R-fng mRNA starts to be detected at stage 15 HH³⁴ in the presumptive dorsal-limb ectoderm (Fig. 1a, b), and is absent from the presumptive ventral ectoderm. At stages 16-18, *R-fng* transcripts are clearly restricted to the dorsal side with a higher concentration at the dorsoventral boundary and later in the AER (Fig. 1d, f). Expression remains restricted to the dorsal ectoderm and the AER until stages 22-24 when it starts to fade in the dorsal ectoderm (Fig. 1h-j), but remains in the AER until the latest stage examined (stage 27). The early expression pattern of chick *R-fng* is similar to that of its *Drosophila* homologue, which is restricted to the dorsal region of the wing imaginal disc²⁵.

In comparison with *R-fng*, expression of *Fgf-8*, a gene thought to play a role in formation of the AER^{23,24}, is not detectable at stage 15

(Fig. 1c), but appears a stage later at the dorsoventral boundary and is subsequently restricted to the AER (Fig. 1e, g).

To explore potential parallels between *Drosophila* and vertebrates, we looked for homologues of downstream genes in the *fringe* pathway (*Delta*, *Serrate* and *Notch*^{26,29,30}). Several vertebrate homologues of each gene were cloned. Of these, one *Serrate* clone and one *Notch* clone were found to be expressed in the AER. This is shown in Fig. 2, along with the expression patterns of *Fgf-8*, *Wnt-7a* and *En-1*.

R-fng regulates AER formation

The role of *R-fng* in vertebrate limb development was studied by inducing its ectopic expression in the ventral portion of the limb. About 300 limb primordia were infected with a replication-competent retroviral vector containing *R-fng*; an abnormal limb phenotype was observed in 16% of the infected embryos.

The limb-bud phenotypes were found to be quite variable, and

can be divided into four classes. In some embryos limb development was arrested after viral infection. These limbs lack a defined AER, which would account for the reduced limb outgrowth (Fig. 3c). In the second phenotype a marked deletion of distal limb tissue results in an AER divided into two or more domains (Fig. 3d, e). A third class of embryos show a phenotype with additional ventral or ventrally expanded AERs giving rise to abnormal outgrowths (Fig. 3f, g); additional or expanded AERs were never observed on the dorsal side of the limb. Finally, some infected embryos exhibited an enlarged limb bud with a lengthwise expansion of the anterior AER (Fig. 3h). All four phenotypes suggest that *R-fng* misexpression perturbs the normal formation of the AER.

When the embryos were allowed to develop further (up to 10 days), the cartilage patterns were found to correlate well with the AER perturbations seen in the early limb buds. In some embryos the limbs were found to be severely truncated, showing a very reduced

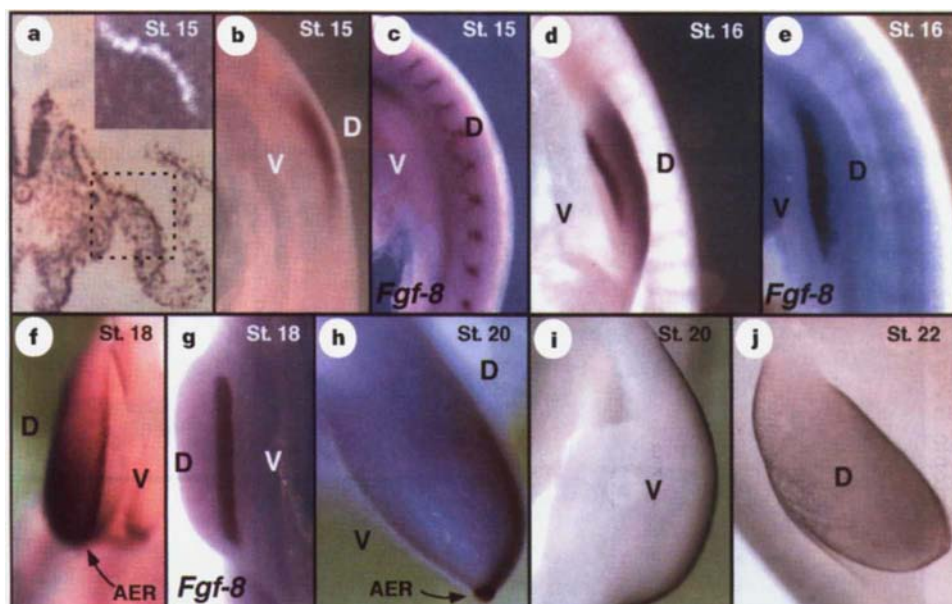


Figure 1 Spatiotemporal pattern of expression of *R-fng* and *Fgf-8* during chick limb-bud development. **a, b**, *R-fng* transcripts start to be detected at stage 15 in the presumptive dorsal (D) limb ectoderm. **c**, At the same stage (st) (15), *Fgf-8* transcripts cannot be detected. **d**, At stage 16, higher levels of *R-fng* transcripts begin to accumulate at the dorsoventral boundary, with expression restricted to the dorsal side of the boundary. **e**, *Fgf-8* transcripts appear at stage 16 in a broad ectodermal band at the dorsoventral limb bud boundary. The expression domain

includes both dorsal and ventral (V) cells. **f**, At stage 18, a sharp boundary of *R-fng* expression is seen between the dorsal and ventral limb ectoderm, in addition to expression in the dorsal ectoderm. **g**, At the same stage, *Fgf-8* expression becomes tightly restricted to the AER. **h-j**, *R-fng* expression in the dorsal side of the limb remains until stages 22-24 when it begins to fade. Expression in the AER remains until at least stage 27.

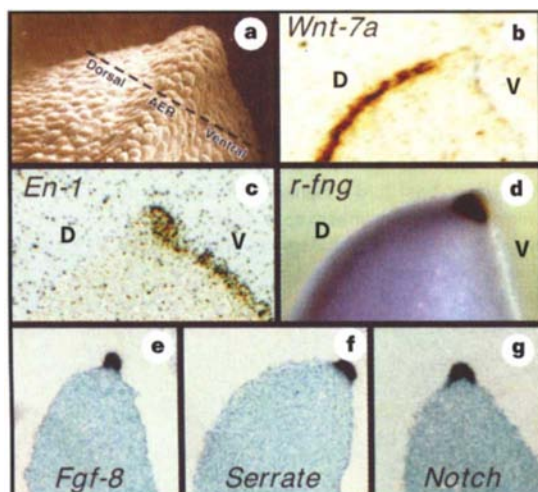


Figure 2 Expression domains of *Wnt-7a*, *En-1*, *R-fng*, *Fgf-8*, *Serrate* and *Notch*. **a**, Scanning electron micrograph of a chicken limb bud (reproduced from ref. 51 (Sinauer) with permission from K. Tosney) illustrating the morphology of the AER and the plane of the sections shown below. **b-g**, Expression domains of *Wnt-7a*, *En-1*, *R-fng*, *Fgf-8*, *Serrate* and *Notch* at stage 20. *Wnt-7a* transcripts are restricted to the dorsal ectoderm. *En-1* expression is restricted to the ventral ectoderm. *R-fng* is expressed in both the dorsal ectoderm and at higher levels in the AER. *Fgf-8* and *Notch* are specifically restricted to the AER; *Serrate* is predominantly restricted to the AER. We have confirmed that the *Notch* clone we show here is identical to that reported previously¹³ and the *Serrate* clone is the chick homologue of the *Serrate-2* gene reported in rat¹⁴. D, dorsal; V, ventral.

and deformed humerus or femur with no distal structures (Fig. 3k, l). In other embryos additional distal structures (most frequently extra digits) were observed (Fig. 3m–p). In most of these cases the proximal structures were normal, but in some there was an accompanying reduction of the proximal elements (Fig. 3n). The number and identity of the extra digits was variable, but most commonly digit 2 was duplicated. The extra digits were either in plane or ventral to the other digits (Fig. 3o, p). Additional structures were never observed on the dorsal side of the limb.

The most likely explanation for the variability of the observed phenotypes is that the extent of viral infection varies in the different embryos. Indeed, limb buds that exhibit arrested growth have uniform viral infection (Fig. 4a), and limb buds with expanded or additional AERs have rather patchy viral infection (Fig. 4b–e). In addition to the variability of the extent of viral infection, it is

possible that the dose is also variable, but this is not easily assessed by *in situ* hybridizations. Furthermore, it is difficult to assess the extent of viral infection at the time the phenotype is triggered as the virus will continue to spread through adjacent tissue. Although the inherent variability of viral infection is commonly considered to be a limitation of this technique, in this case the observation of different phenotypes has greatly assisted our interpretation, and seems to indicate that the phenotype is highly dependent on the pattern of *R-fng* misexpression.

To confirm that the AER-like structures formed as a result of *fng* misexpression were indeed ectopic AERs, and competent to follow normal developmental programs, we characterized the expression of several molecular markers. *Fgf-8* is normally expressed in the AER; several lines of evidence indicate that it is involved in limb initiation as well as maintaining limb outgrowth^{23,24}. Because the

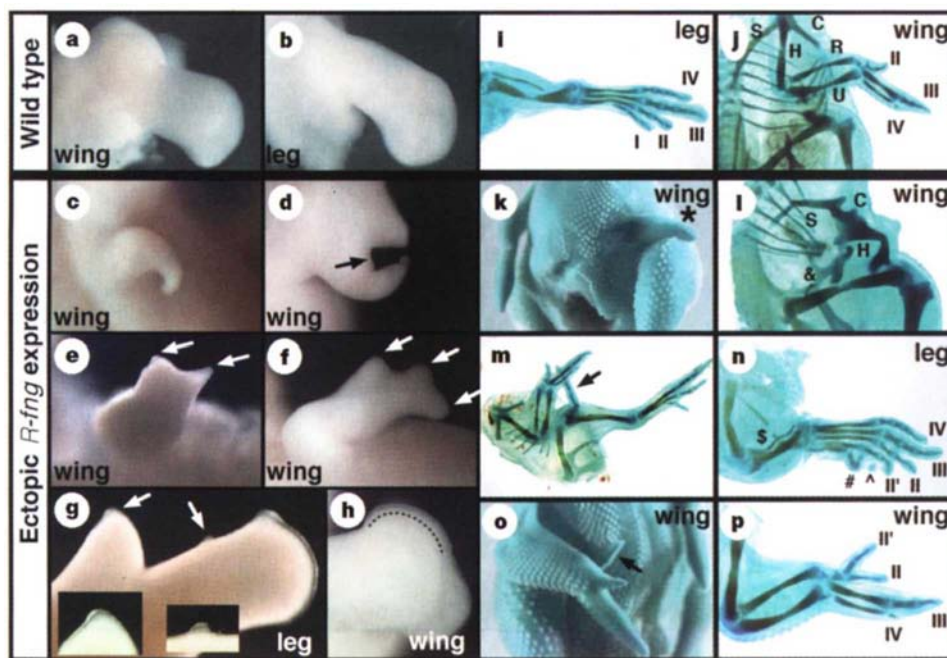


Figure 3 Ectopic *R-fng* expression perturbs normal limb outgrowth. **a, b**, Wild-type limb buds at stage 25. **c–h**, Limb-bud phenotypes are shown at stages 24–25 following injection with virus expressing *R-fng* at stages 8–10. **c**, An extremely reduced wing bud. Note the pointed shape and the absence of a defined AER. **d**, A wing bud showing a marked deletion of tissue (arrow) at its distal end. **e**, A wing bud with a similar, but less severe, deletion of distal tissue leading to an apparent bifurcation of the normal AER (arrows). **f, g**, Limb buds exhibiting additional ventral AERs. The ventral nature of these outgrowths is not easy to visualize (see also **m** and **o**). **h**, Wing bud showing a significant anterior expansion of the AER

(broken line). **i, j**, Wild-type, and **k–p**, *R-fng*-infected limbs 8 days after injection at stages 8–10. **k, l**, Severely reduced wing (asterisk) with deformed scapula (S), humerus (H) and coracoid (C), and absence of radius, ulna and digits; also, additional abnormal skeletal elements (&). **m**, Leg with an additional digit (IV) attached ventrally at the junction between femur and tibia. **n**, Leg with an additional digit II (II'), an unidentified structure (A) and a deformed digit II (#). The tibia and fibula appear shortened and deformed (\$). **o, p**, Wing with additional digit II (arrow and II') arising from the ventral side of the normal limb, although this is clearer in **o** as the embryo was flattened in **p** to facilitate photography.

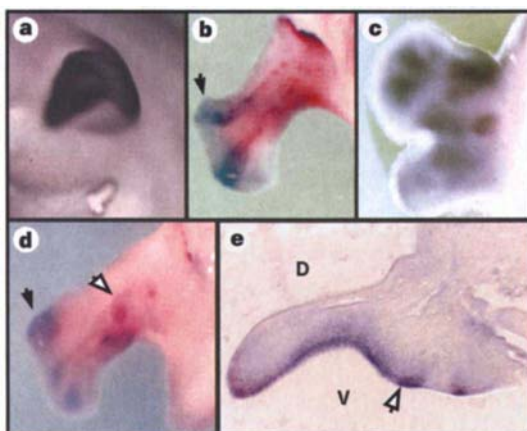


Figure 4 Correlation of viral infection with phenotype. **a**, Truncated wing bud 55 h after viral infection. The uniform viral infection correlates with the severe truncation. **b–d**, Ventral views of enlarged limb buds at stages 25–26. The limbs show patchy viral infection. In **b** and **d** these patches correlate with anterior outgrowth (filled arrows). In **d** a marked ventral outgrowth (hollow arrow) also shows a region of viral infection at its tip. **e**, A section from a significantly reduced limb that has a large region of virally infected tissue on the ventral side. The hollow arrow indicates a ventral outgrowth located at a boundary of the viral infection and hence *R-fng* expression.

misexpression of *R-fng* can lead either to the disruption of the AER (hence arrested limb outgrowth) or to the formation of additional or expanded AERs (hence additional limb outgrowth), we would expect to observe perturbations in the *Fgf-8* expression pattern. *In situ* hybridization 55–60 h after viral infection shows that *Fgf-8* expression is absent in the reduced limb buds (Fig. 5a). In limb buds with additional or ventrally expanded AERs, *Fgf-8* transcripts were found to colocalize with the additional AERs (Fig. 5b–e).

The appearance of additional skeletal elements after *R-fng* misexpression would suggest that genes involved in the establishment of the anteroposterior axis are dependent on *R-fng* expression^{7,17,21,39}. Consistent with this, there are enlarged areas of expression of *hoxd-12* and *Fgf-4* in the anterior mesoderm and AER, respectively (Fig. 5f, g). In a few cases, ectopic patches of *Shh* expression were observed at the anterior margin of *R-fng*-infected limb buds (data not shown).

To determine whether *R-fng* influences dorsoventral patterning we examined the expression of *Lmx-1*, an established marker for dorsal mesoderm identity^{13,14}. Because *Lmx-1* expression is absent from both sides of the ventrally derived outgrowths, it seems that these are biventral (data not shown).

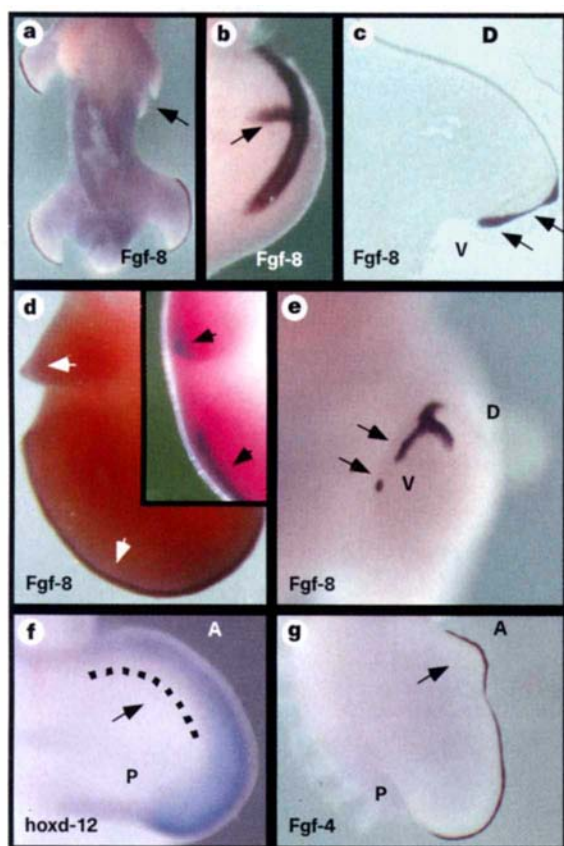


Figure 5 Misexpression of *R-fng* perturbs the normal expression patterns of *Fgf-8*, *hoxd-12* and *Fgf-4* genes. **a**, *Fgf-8* transcripts are absent from truncated limb buds that result from extensive ectopic *R-fng* expression. **b–e**, Arrows indicate ectopic ventral expression of *Fgf-8* in the additional AERs induced by *R-fng* misexpression. **b**, An ectopic stripe of *Fgf-8* on the ventral side of the *R-fng*-infected limb bud. **c**, *Fgf-8* expression in a section through a ventrally expanded AER. **d**, **e**, Whole-mount *in situ* hybridizations with *Fgf-8* expression in ectopic AERs that are either discrete (**d**) or intersect (**e**) the normal AER. **f**, *Hox-d12* transcripts, normally restricted to the posterior-distal limb-bud mesoderm, are induced at the anterior-distal mesoderm after *R-fng* viral infection (broken line). **g**, Similarly, *Fgf-4*, restricted to the posterior region of the AER during normal limb development, is induced in the anteriorly expanded AER after *R-fng* viral infection (arrow).

Influence of *Wnt-7a* and *En-1* on *R-fng* expression

Although *R-fng* does not affect dorsoventral identity, several lines of evidence suggest that *R-fng* expression is dependent on previously established dorsoventral cues: (1) *R-fng* expression is restricted to the dorsal ectoderm and directs AER formation at the dorsoventral boundary; (2) correct AER positioning depends on correct *En-1* expression because in mice lacking *En-1*, the AER is significantly expanded on the ventral side of the limb bud¹⁵; (3) in the chick *limbless* mutant, no AER is formed, *En-1* is absent, and *Wnt-7a* is expressed uniformly in the limb bud ectoderm³⁵; and (4) in *Drosophila*, the dorsal selector gene *apterous* regulates *R-fng* expression²⁵. To determine whether either of the ectodermal genes known to specify dorsoventral identity (*Wnt-7a* and *En-1*) influences *R-fng* expression, we used the viral misexpression technology.

As previously reported¹⁴, limbs infected with virus expressing *Wnt-7a* show a bidorsal character, but in no case was any perturbation of the pattern of *R-fng* expression observed (Fig. 6a–c).

Approximately half of the embryos ($n = 58$) infected with virus expressing *En-1* show striking perturbations of limb outgrowth. In some cases a severe truncation of the limb bud was observed, the AER was not well defined, and the limbs exhibited a ‘rippled’

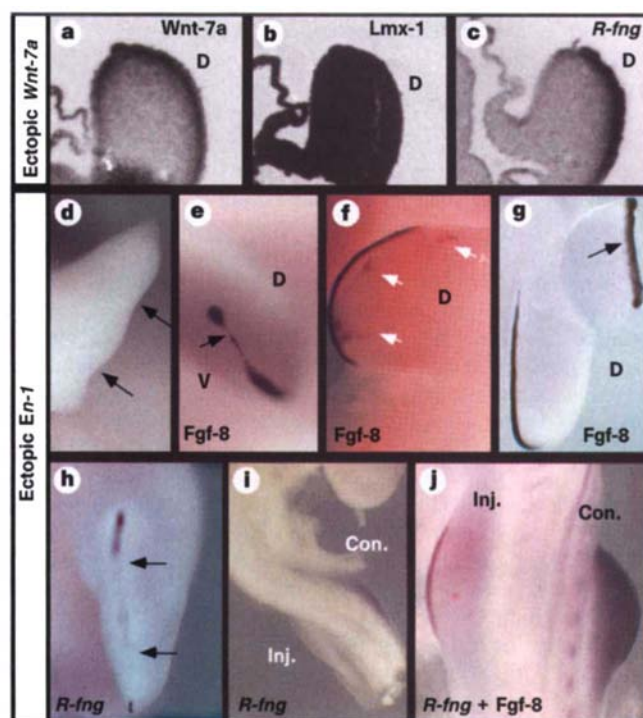


Figure 6 *Wnt-7a* and *En-1* misexpression. **a**, 48 h after *Wnt-7a* viral infection of the limb primordia at stage 12, ectopic patches of *Wnt-7a* are detected in the ventral ectoderm and AER as well as in proximal regions of the limb bud. **b**, Ectopic expression of *Wnt-7a* on the ventral side of the limb induces ectopic ventral expression of *Lmx-1*, but **c**, it does not perturb the expression of *R-fng*. **d–h**, Limb buds fixed 55 h after *En-1* viral infection at stages 8–10. In some cases (**d**, **e**, **h**) the limbs show a rippled or discontinuous AERs; in others (**f**, **g**) ectopic dorsal outgrowths appear (arrows). This correlates with discontinuous (**e**) or ectopic dorsal expression of *Fgf-8* (**f**, **g**). **h–j**, *En-1* misexpression represses *R-fng* in the AER and dorsal ectoderm of infected limbs. Arrows in **h** indicate discontinuities of *R-fng* expression in the AER. **i**, Embryo fixed at stage 17 after *En-1* misexpression. The injected bud (Inj.) shows a clear reduction in *R-fng* expression when compared with the contralateral one (Con.). **j**, The injected *En-1* embryo was fixed at stage 18 and hybridized simultaneously with *R-fng* and *Fgf-8* probes. Although some staining of *Fgf-8*/*R-fng* is still observed in the AER of the injected limb (Inj.), the dorsal ectoderm expression of *R-fng* is greatly reduced.

appearance with an apparently discontinuous AER (Fig. 6d). In other cases ectopic outgrowths or additional AERs were observed on the dorsal side of the normal limb (Fig. 6f, g). Whole-mount *in situ* hybridization at various stages using probes for *Fgf-8* indicate discontinuous expression in the rippled AERs (Fig. 6e) and ectopic expression in the additional AERs of the dorsal outgrowths (Fig. 6f, g). Hybridization using probes for *Wnt-7a*, *Lmx-1* (data not shown) and *R-fng* show that misexpression of *En-1* strikingly represses expression of these genes. Where viral infection is extensive (data not shown) this correlates with truncated limb buds that exhibit no *R-fng* expression (Fig. 6i, j). Where breaks in the normal AERs are observed, the *R-fng* expression was found to be discontinuous (Fig. 6h).

Although the phenotypes of *En-1* misexpression are varied, they are all consistent with the proposal that ectopic *En-1* expression on the dorsal side of the limb-bud represses *r-fng* expression. Thus it seems likely that a deformed structure results when *R-fng* expression is repressed in the presumptive AER, and additional ectopic AERs are induced when *En-1* misexpression causes non-uniform *R-fng* expression in the dorsal ectoderm. These data, taken together with the knockouts of *Wnt-7a* and *En-1*, strongly suggest that a correct pattern of *En-1* expression, but maybe not *Wnt-7a* expression, is necessary for the formation and positioning of the AER.

fng activity mediates boundary interactions

Not all of the genes involved in the control of limb outgrowth and dorsoventral patterning have yet been identified, but we are beginning to understand how the interactions between the known genes combine to regulate limb development. Our experiments suggest that the AER develops at the interface of tissue that is expressing high levels of *R-fng* adjacent to tissue that does not express *R-fng*, that is, at a *R-fng* boundary. Because *R-fng* is not normally expressed on the ventral side of the limb, ectopic expression of *R-fng* here creates new boundaries that result in additional AERs. Expression of *En-1* on the dorsal side represses normal *R-fng* expression and can therefore create new boundaries and ectopic outgrowths. Finally, when *R-fng* is expressed uniformly on both sides of the limb, or when it is absent from both sides of the limb owing to repression by ectopic *En-1* expression, then no *R-fng* boundaries are formed, the AER does not develop, and growth is arrested. It is important to note that in the misexpression experiments the ectopic outgrowths always have a character consistent with their dorsoventral location, indicating that *R-fng* does not influence dorsoventral identity.

The observation that misexpression of *En-1* in the dorsal ectoderm leads to areas without *r-fng* transcripts indicates that *En-1* directly or indirectly represses *R-fng* expression. It is therefore possible that repression by *En-1* defines the dorsal expression domain of *R-fng* when it first appears at stage 15. This suggests that *En-1* may play a critical role in positioning the AER at the dorsoventral boundary. This is supported by the observation that in the early limb buds of the chick *limbless* mutant, *En-1* expression is absent³⁵, *R-fng* is expressed in both the dorsal and ventral ectoderm³⁶, and no AER is formed. The importance of *En-1* in positioning the AER is consistent with the finding that, in mice lacking *En-1*, the AER is mispositioned and expanded ventrally. However, that an AER is formed at all suggests that there is a ventrally displaced *R-fng* boundary, implying that other factors contribute to restricting the domain of *R-fng* expression. It will be important to examine the pattern of *R-fng* expression in mice lacking *En-1*, and also to explain why *R-fng* and *En-1* expression overlap in the ventral side of the mature AER (Fig. 2).

In *Drosophila*, gain- and loss-of-function experiments indicate that wing-margin formation and wing outgrowth occur at the interface of cells expressing *fng* and cells not expressing *fng*²⁵. This is similar to our observation in vertebrates. Together these results indicate the existence of a common signalling process mediated by *fng* in the formation of both the vertebrate AER and the insect wing margin.

We can envisage several mechanisms through which *fng* boundaries might generate an environment that could initiate AER formation. A straightforward model proposed for *Drosophila*²⁵ is that, in those cells expressing *fng*, *fng* receptors are repressed by the ligand itself. Considering that *R-fng* may be a secreted glycosyltransferase, the 'receptor' could be an extracellular carbohydrate substrate, bound to a protein or lipid, whose activity is modulated by *R-fng*. We do not yet know what this target is, but candidates include other extracellular signalling molecules and their receptors. Whatever the nature of the *R-fng* target, it seems either that only ventral cells (lacking *R-fng* expression) are able to respond or that only dorsal cells adjacent to ventral cells can respond. This could be because the *R-fng* target is absent or inactive in *R-fng*-expressing cells on the dorsal side of the limb bud, or because the target is specific to ventral cells, perhaps as a consequence of *En-1* expression. Further insight will require more experiments making use of knockout mice, viral misexpression techniques, biochemical and fate map studies.

A simplified scheme for AER formation is shown in Fig. 7. At stage 14 interaction between the mesoderm and ectoderm establishes the dorsoventral boundary in the ectoderm and the asymmetric pattern of *R-fng* expression. After the *R-fng*-mediated signalling event, cells at the dorsoventral boundary presumably respond and initiate AER formation, perhaps by further signalling to adjacent cells or by establishing new tissue by cell division. Once *R-fng* signalling is established, downstream genes are activated and drive subsequent limb outgrowth. A candidate gene active early in this pathway, or perhaps cooperating with *R-fng*, is *Fgf-8*. We have shown (Fig. 1) that transcripts for *Fgf-8* are detected slightly later than *R-fng* mRNA, just before the formation of a morphologically

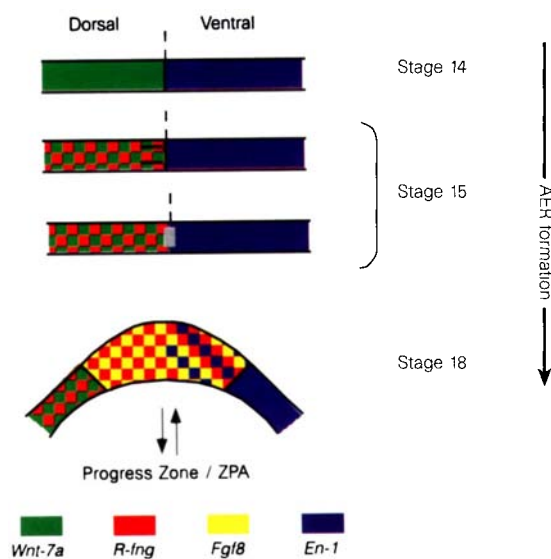


Figure 7 *R-fng* signalling and AER formation. A schematic illustration of the expression patterns of genes present in the ectoderm at the dorsoventral boundary (broken line) during AER formation. At stage 14, expression of *Wnt-7a* and *En-1* is restricted to the dorsal and ventral sides of the presumptive limb ectoderm, respectively. At stage 15, *R-fng* expression appears in the dorsal side of the ectoderm and forms a *R-fng*/no-*R-fng* boundary. It is likely that *R-fng* signals (arrows) to cells immediately at the dorsoventral boundary. (Note that this model suggests that the primary role of *R-fng* in AER formation occurs before stage 17.) These cells then give rise to the AER, either as a consequence of further signalling to adjacent cells or by cell division. In the mature AER, two cell types can be distinguished characterized by *R-fng*/*Fgf-8* expression with no *En-1* or *Wnt-7a*; *R-fng*/*Fgf-8*/*En-1* expression. The layout of the figure suggests that the dorsal half of the AER derives from dorsal cells, and the ventral half from ventral cells, but this need not be the case and requires further investigation. Arrows beneath the ectoderm indicate that there is an interplay with signals from the underlying mesoderm (progress zone and polarizing region, ZPA).

distinguishable AER, in a symmetrical stripe of cells encompassing the dorsoventral ectodermal boundary. If *Fgf-8* is indeed involved at an early stage in the *fng/no-fng* effector pathway then we should expect that its misexpression would have a similar phenotype to creating ectopic *fng/no-fng* boundaries. Indeed, application of *Fgf-8* to the early limb bud of the *limbless* mutant can rescue the phenotype and give rise to limb outgrowth³⁵.

Evolutionary conservation of developmental programs

The observation that *fng* homologues in vertebrates seem to operate in essentially the same way as their counterparts in *Drosophila* raises the general question of the similarity of limb development in these distant organisms. The insect and vertebrate limbs are morphologically quite distinct, and so it is difficult to correlate particular tissues and structures in the two species. However, it is becoming clear that both limbs develop through the integration of different signals from the dorsoventral and anteroposterior axes that allow further elaboration of the proximodistal axis and outgrowth of the limb^{11,12,17,21,27,28,37}. Remarkably, some of the molecules involved in this process seem to be conserved. In particular, *hedgehog*, *patched* and *decapentaplegic* in *Drosophila* and their counterparts in vertebrates (*Sonic hedgehog*, *patched* and the *Bone morphogenetic proteins*) establish the anteroposterior axis in both species^{21,38–41}. In contrast to the similarities along the anteroposterior limb axis in insects and vertebrates, conservation of the molecules that govern dorsoventral identity is less apparent. *Lmx-1* has been suggested to be a vertebrate counterpart of *apterous*, but here we show that, unlike *apterous*, it is not involved in *R-fng* expression. *Wnt-7a* does not regulate *R-fng* expression, and does not seem to have a functional counterpart in the dorsal compartment of the *Drosophila* wing, and the homologue of *En-1* in *Drosophila* is involved in establishing the anteroposterior axis⁴². However, with the characterization of chick *fng*, we have shown that a key player in the transduction of dorsoventral signalling to proximodistal signalling is highly conserved between insects and vertebrates. Because chick homologues of the genes *Serrate* and *Notch*^{43,44} are also expressed in the AER, it seems that in general, the genes responsible for transducing the dorsoventral information at the boundary, and hence establishing the proximodistal pathway, are conserved, although functional equivalence has yet to be demonstrated. Interestingly, counterparts of other genes known to be involved in the formation of AER and wing margin (the *Fgf* genes in vertebrates and *vestigial* in *Drosophila*) have not yet been identified.

The apparent conservation of *R-fng* and downstream molecules in the proximodistal pathway suggests that these comprise a conserved developmental unit, or syntagma⁴⁵. Some of the genes that direct the formation of the anteroposterior axis form a similarly conserved syntagma. The conservation of these two developmental subroutines highlights the apparent lack of conservation in the molecules that establish the dorsoventral axis. It will be important to establish whether these apparent differences are real differences responsible for establishing evolutionary diversity between insect and vertebrate limbs, or whether homologous counterparts will be discovered. In conclusion, our results suggest that, in making a limb, a similar combination of genetic building blocks or syntagmata have been retained in flies and vertebrates, although the identity of certain 'selector' genes may have changed. These changes may in turn subtly alter the interaction between the different syntagmata accounting for differences in the morphology of insect and vertebrate limbs. Nevertheless, the basic functional homologues, such as the vertebrate AER and the *Drosophila* wing margin, have been preserved. □

Methods

Identification of *R-fng*. The probe used for screening covered the entire open reading frame of *Drosophila fng* and was provided by K. Irvine and E. Wieschaus. Approximately 2×10^6 phage from a chick limb cDNA library

(stages 20–23; ref. 34) were screened under low-stringency conditions⁴⁶. Three of these were found to encode *R-fng*. Several related clones were also obtained, and will be published elsewhere. Positive clones were excised and sequenced according to manufacturer's instructions (Stratagene, Qiagen and USB).

In situ hybridization. Whole-mount and sectional *in situ* hybridizations using digoxigenin and ³⁵S-labelled probes were performed as described previously⁴⁷. The *r-fng* digoxigenin and ³⁵S-labelled antisense probes span the entire open reading frame. The *Serrate-2* and *Notch* antisense probes span the intracellular domain. The details of the remaining probes have been reported elsewhere: *virus* (ref. 21); *Fgf-8* (ref. 24); *Hoxd-12* (ref. 48); *Fgf-4* (ref. 16); *Lmx-1* and *Wnt-7a* (ref. 14).

Retroviral infection. Chicken embryos (either MacIntyre Poultry (San Diego, CA) or SPAFAS (Norwich, CT)) were infected with virus in the wing or leg primordia at stages 8–10 as previously described⁴⁹. The replication-competent retroviral vector used, RCAS(BP)A⁵⁰, contained cDNAs spanning the entire coding region of the *Wnt-7a* (provided by A. MacMahon and L. Niswander), *R-fng*, or *En-1* genes, respectively. Virus preparation was carried out as described previously¹⁴. After different incubation time points, embryos were fixed in 4% paraformaldehyde overnight and dehydrated in methanol. After evaluation under a dissecting microscope for possible alterations, embryos were stored at –20 °C until processed for *in situ* hybridization. Visualization of cartilage structure was as previously described¹⁴.

Received 20 December 1996; accepted 6 March 1997.

- Saunders, J. W. & Gasseling, M. T. in *Epithelial-Mesenchymal Interactions* (eds Fleischmajer, R. & Billingham, R. E.) 78–97 (Baltimore, 1968).
- Tickle, C., Summerbell, D. & Wolpert, L. Positional signalling and specification of digits in chick limb morphogenesis. *Nature* **254**, 199–202 (1975).
- Pautou, M. P. & Kieny, M. Interaction ec-mesodermique dans l'établissement de la polarité dorso-ventrale du pied de l'embryon de poulet. *C. r. hebdomadaire Séances Acad. Sci., Paris D* **277**, 1225–1228 (1973).
- MacCabe, J. A., Erick, J. E. & Saunders, J. W. J. Ectodermal control of the dorso-ventral axis in the leg bud of the chick embryo. *Dev. Biol.* **39**, 69–82 (1974).
- Saunders, J. W. J. The proximo-distal sequence of the origin of the parts of the chick wing and the role of the ectoderm. *J. Exp. Zool.* **108**, 363–404 (1948).
- Todt, W. & Fallon, J. F. Development of the apical ectodermal ridge in the chick wing bud. *J. Embryol. Exp. Morphol.* **80**, 21–41 (1984).
- Duboule, D. How to make a limb? *Science* **266**, 575–576 (1994).
- Maden, M. The limb bud—part two. *Nature* **371**, 560–561 (1994).
- Martin, G. Why thumbs are up. *Nature* **374**, 410–411 (1995).
- Tabin, C. The initiation of the limb bud: Growth factors, *Hox* genes, and retinoids. *Cell* **80**, 671–674 (1995).
- Parr, B. A. & McMahon, A. P. Dorsalizing signal *Wnt-7a* required for normal polarity of D–V and A–P axes of mouse limb. *Nature* **374**, 350–353 (1995).
- Yang, Y. Z. & Niswander, L. Interaction between the signaling molecules *Wnt7a* and *shh* during vertebrate limb development—dorsal signals regulated anteroposterior patterning. *Cell* **80**, 939–947 (1995).
- Riddle, R. et al. Induction of the LIM Homeobox gene *Lmx-1* by *Wnt7a* establishes dorsoventral pattern in the vertebrate limb. *Cell* **83**, 631–640 (1995).
- Vogel, A., Rodriguez, C., Warnken, W. & Izpisua Belmonte, J. C. Dorsal cell fate specified by chick *Lmx1* during vertebrate limb development. *Nature* **378**, 716–720 (1995).
- Loomis, C. A. et al. The mouse *Engrailed-1* gene and ventral limb patterning. *Nature* **382**, 360–363 (1996).
- Niswander, L., Tickle, C., Vogel, A., Booth, I. & Martin, G. R. FGF-4 replaces the apical ectodermal ridge and directs outgrowth and patterning of the limb. *Cell* **75**, 579–587 (1993).
- Niswander, L., Jeffrey, S., Martin, G. R. & Tickle, C. A. A positive feedback loop coordinates growth and patterning in the vertebrate limb. *Nature* **371**, 609–612 (1994).
- Fallon, J. F. et al. FGF-2: Apical ectodermal ridge growth signal for chick limb development. *Science* **264**, 104–107 (1994).
- Ohuchi, H. et al. An additional limb can be induced from the flank of the chick embryo by FGF4. *Biochem. Biophys. Res. Commun.* **209**, 809–816 (1995).
- Cohn, M., Izpisua Belmonte, J. C., Abud, H., Heath, J. K. & Tickle, C. FGF-2 application can induce additional limb formation from the flank of chick embryos. *Cell* **80**, 739–746 (1995).
- Laufer, E., Nelson, C. E., Johnson, R. L., Morgan, B. A. & Tabin, C. *Sonic hedgehog* and *Fgf-4* act through a signalling cascade and feedback loop to integrate growth and patterning of the developing limb bud. *Cell* **79**, 993–1003 (1994).
- Mahmood, R. et al. A role for FGF8 in the initiation and maintenance of vertebrate limb bud outgrowth. *Curr. Biol.* **5**, 797–806 (1995).
- Crossley, P. H., Minowada, G., MacArthur, C. A. & Martin, G. R. Roles for FGF-8 in the induction, initiation and maintenance of chick limb development. *Cell* **84**, 127–136 (1996).
- Vogel, A., Rodriguez, C. & Izpisua Belmonte, J. C. Involvement of FGF-8 in initiation, outgrowth and patterning of the vertebrate limb. *Development* **122**, 1737–1750 (1996).
- Irvine, K. D. & Wieschaus, E. *Fringe*, a boundary-specific signaling molecule, mediates interactions between dorsal and ventral cells during *Drosophila* wing development. *Cell* **79**, 595–606 (1994).
- Kim, J., Irvine, K. D. & Carroll, S. B. Cell recognition, signal induction, and symmetrical gene activation at the dorsal-ventral boundary of the developing *Drosophila* wing. *Cell* **82**, 795–802 (1995).
- Diaz-Benjumea, F. J. & Cohen, S. M. *Serrate* signals through *Notch* to establish a *Wingless*-dependent organizer at the dorsal/ventral compartment boundary of the *Drosophila* wing. *Development* **121**, 4215–4225 (1995).
- Kim, J. et al. Integration of positional signals and regulation of wing formation and identity by *Drosophila vestigial* gene. *Nature* **382**, 133–138 (1996).
- Neumann, C. J. & Cohen, S. M. A hierarchy of cross-regulation involving *Notch*, *wingless*, *vestigial* and *cut* organizes the dorsal-ventral axis of the *Drosophila* wing. *Development* **122**, 3477–3485 (1996).
- De Celis, J. F., Garcia-Bellido, A. & Bray, S. J. Activation and function of *Notch* at the dorsal-ventral boundary of the wing imaginal disc. *Development* **122**, 359–369 (1996).
- Wu, J. Y., Wen, L., Zhang, W.-J. & Rao, Y. The secreted product of *Xenopus* gene *lunatic fringe*, a vertebrate signaling molecule. *Science* **273**, 355–358 (1996).

32. Yuan, Y. P., Schultz, J., Mlodzik, M. & Bork, P. Secreted *Fringe*-like signalling molecules may be glycosyltransferases. *Cell* **88**, 9–11 (1997).
33. Cope, L. D. *et al.* Molecular cloning of a gene involved in lipopoligosaccharide biosynthesis and virulence expression by *Haemophilus influenzae* type B. *Mol. Microbiol.* **5**, 1113–1124 (1991).
34. Hamburger, V. & Hamilton, H. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–92 (1951).
35. Ros, M. A. *et al.* The limb field mesoderm determines initial limb bud anteroposterior asymmetry and budding independent of *sonic hedgehog* or apical ectodermal gene expressions. *Development* **122**, 2319–2330 (1996).
36. Laufer, E. *et al.* Expression of *Radical fringe* in limb-bud ectoderm regulates apical ectodermal ridge formation. *Nature* **386**, 366–373 (1997).
37. Couso, J. P., Kust, E. & Martinez Arias, A. *Serrate* and *wingless* in *Drosophila* wing development. *Curr. Biol.* **5**, 1437–1448 (1995).
38. Goodrich, L. V., Johnson, R. L., Milenkovic, L., McMahon, J. A. & Scott, M. P. Conservation of the *Hedgehog/patched*/signaling pathway from flies to mice: Induction of a mouse *patched* gene by *Hedgehog*. *Genes Dev.* **10**, 301–312 (1996).
39. Riddle, R. D., Johnson, R. L., Laufer, E. & Tabin, C. *Sonic hedgehog* mediates the polarizing activity of the ZPA. *Cell* **75**, 1401–1416 (1993).
40. Francis, P. H., Richardson, M. K., Brickell, P. M. & Tickle, C. Bone morphogenetic proteins and a signaling pathway that controls pattern in the developing chick limb. *Development* **120**, 209–218 (1994).
41. Marigo, V., Scott, M. P., Johnson, R. L., Goodrich, L. V. & Tabin, C. J. Conservation in *hedgehog* signaling: induction of a chicken *patched* homolog by *sonic hedgehog* in the developing limb. *Development* **122**, 1225–1233 (1996).
42. Zecca, M., Basler, K. & Struhl, G. Sequential organizing activities of *engrailed* *hedgehog* and decapentaplegic in the *Drosophila* wing. *Development* **121**, 2265–2278 (1995).
43. Myat, A., Henrique, D., Ish-Horowitz, D. & Lewis, J. A chick homologue of *serrate* and its relationship with *notch* and *delta* homologues during central neurogenesis. *Dev. Biol.* **174**, 233–247 (1996).
44. Shawber, C. J., Boulter, J. & Weinmaster, G. Jagged2: a *serrate*-like gene expressed during rat embryogenesis. *Dev. Biol.* **180**, 370–376 (1996).
45. Garcia-Bellido, A. How organisms are put together. *Eur. Rev.* **2**, 15–21 (1994).
46. Izpisua Belmonte, J. C., Falkenstein, H., Dolle, P., Renucci, A. & Duboule, D. Murine genes related to the *Drosophila* *AbdB* homeotic gene are sequentially expressed during development of the posterior part of the body. *EMBO J.* **10**, 2279–2289 (1991).
47. Izpisua Belmonte, J. C., De Robertis, E. M., Storey, K. G. & Stern, C. The homeobox gene *gooseoid* and the origin of organizer cells in the early chick blastoderm. *Cell* **74**, 645–659 (1993).
48. Izpisua Belmonte, J. C., Tickle, C., Dolle, P., Wolpert, L. & Duboule, D. Expression of the homeobox *Hox-4* genes and the specification of position in chick wing development. *Nature* **350**, 585–589 (1991).
49. Morgan, B. A., Izpisua Belmonte, J. C., Duboule, D. & Tabin, C. J. Targeted misexpression of *Hox-4.6* in the avian limb bud causes apparent homeotic transformations. *Nature* **358**, 236–239 (1992).
50. Hughes, S., Greenhouse, J., Petropoulos, J. & Sutcliffe, P. Adaptor plasmids simplify the insertion of foreign DNA into helper-independent retroviral vectors. *J. Virol.* **61**, 3004–3012 (1987).
51. Gilbert, S. F. *Dev. Biol.* (Sinauer, Sunderland, MA, 1995).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank C. Tabin, J. Fallon and members of their laboratories for sharing information and ideas before publication; C. Kintner and R. Evans for discussions; B. Blumberg, S. Bryant, A. Vogel, A. Tavares, H. Juguilon, A. Tugores, J. Cleary and C. De La Peña for comments on the manuscript and technical advice; L. Hooks for assembling the manuscript; and L. Niswander, A. MacMahon, K. Irvine, E. Wieschaus, D. Enrique, A. Joyner, H. Hughes and H. Nakamura for probes and reagents. This work was supported by a HFSPO fellowship to J.W.R.S. National Science Foundation undergraduate fellowships to J.D.L.P. and B.E., and grants from the National Science Foundation, March of Dimes, the G. Harold and Leila Y. Mathers Charitable Foundation, and Kieckhefer Foundation. J.C.I.B. is a Pew scholar.

Correspondence and requests for materials should be addressed to J.C.I.B. (e-mail: carlos_belmonte@qm.salk.edu). The sequence for *R-fng* has been deposited in the Genbank database, accession no. U82088.

Expression of *Radical fringe* in limb-bud ectoderm regulates apical ectodermal ridge formation

Ed Laufer*, Randall Dahn†, Olivia E. Orozco*, Chang-Yeol Yeo*, Jacqueline Pisenti‡, Domingos Henrique§, Ursula K. Abbott‡, John F. Fallon† & Cliff Tabin*

* Department of Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA

† Department of Anatomy, University of Wisconsin, 1300 University Avenue, Madison, Wisconsin 53706, USA

‡ Department of Avian Sciences, University of California, Davis, California 95616, USA

§ Developmental Genetics Laboratory, Imperial Cancer Research Fund, PO Box 123, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

The apical ectodermal ridge of the vertebrate limb bud lies at the junction of the dorsal and ventral ectoderm and directs patterning of the growing limb. Its formation is directed by the boundary between cells that do and cells that do not express the gene *Radical fringe*. This is similar to the establishment of the margin cells at the *Drosophila* wing dorsoventral border by *fringe*. *Radical fringe* expression in chick-limb dorsal ectoderm is established in part through repression by *Engrailed-1* in the ventral ectoderm.

Inductive signals from discrete signalling centres regulate much of embryonic development. The apical ectodermal ridge (AER) is an inductive centre that runs anteroposteriorly along the distal margin of the limb-bud ectoderm, and has several distinct roles during limb ontogeny¹. It maintains developmental plasticity within the underlying limb mesenchyme; maintains expression of the signalling molecule *Sonic hedgehog* in the posterior mesenchyme; synergizes with other factors to pattern the limb mesenchyme; and controls the anteroposterior width of the developing limb bud. The signals produced by the AER that mediate each of these activities are members of the fibroblast growth factor family, notably *Fgf-2*, *Fgf-4* and *Fgf-8*.

The correct spatial and temporal control of the formation of signalling centres such as the AER is critical for normal embryogenesis. However, as with most embryonic signalling centres, its induction is only partly understood. Expression of molecular markers, notably *Fgf-8*, indicates that AER induction begins by stage 16 (refs 2–4). Tissue grafting experiments have shown that

signals from the limb mesoderm induce AER formation in the overlying ectoderm^{5–7}, and that the capacities of both the mesenchyme to signal and the ectoderm to respond diminish over time, being lost from both tissues by stage 20 (ref. 8). Although mesodermal signals trigger AER formation, the information specifying both the position and number of AERs is encoded within the ectoderm^{9,10}.

The AER is positioned at the junction of the dorsal and ventral limb ectoderm. Several genes are expressed in dorsoventral restricted domains within the limb bud before AER formation, and are therefore candidates for regulating AER formation. *Wnt-7a* and *Lmx-1* are expressed in dorsal ectoderm and mesoderm, respectively^{11–13}. They specify dorsal-limb mesodermal fates, but do not seem to be directly involved in AER formation, as neither loss of *Wnt-7a* function¹⁴ nor gain of *Wnt-7a* or *Lmx-1* function^{12,13} alters normal AER formation. *Engrailed-1* (*En-1*) is a ventral limb ectodermal marker that is also expressed in the ventral half of the AER¹⁵. It has been directly implicated in regulating AER formation

or maintenance, as well as in controlling ventral cell fates, as limb buds of *en-1*^{-/-} mice have ventrally extended and flattened AERs and bidorsal distal structures¹⁶.

Two recessive mutant chick strains, *limbless*^{17,18} and *eudiplopodia*¹⁹, are defective in AER regulation, and provide some insights into the mechanisms of AER positioning. In *limbless* homozygote embryos, limb buds initiate normally but regress because they never form an AER. Recombination experiments have shown that it is the *limbless* ectoderm that is defective²⁰. Molecular analysis has since found that *limbless* ectoderm is uniformly dorsal. The dorsal marker *Wnt-7a* is ubiquitously expressed, the ventral marker *En-1* is absent, and rescue of *limbless* limb outgrowth with Fgf-soaked beads results in bidorsal limbs²¹⁻²³. These results have led to the hypothesis that the dorsoventral and AER induction pathways are linked, and that AERs form as a result of interactions between dorsally and ventrally specified tissue; grafting experiments also support this view²⁴. However, limbs of *eudiplopodia* homozygote embryos display a phenotype inconsistent with this model. After the normal AER has formed and limb-bud elongation begins, ectopic AERs form on the dorsal surface of *eudiplopodia* limb buds, giving rise to polydactylous limbs. The

resulting extra digits have a morphologically bidorsal phenotype, suggesting that AERs can arise independently of dorsoventral ectodermal identity. However, molecular analysis of *eudiplopodia* limbs has not been performed, and alternative explanations are plausible.

There are some intriguing functional parallels between the AER of the vertebrate limb bud and the margin cells of the *Drosophila* wing imaginal disc, suggesting that they may be induced by similar mechanisms. Like the AER, the wing margin is established at a dorsoventral boundary and is required for proximodistal wing outgrowth and patterning²⁵. Cells arising from the dorsal half of the margin have dorsal identities, whereas those in the ventral half have ventral identities^{25,26}. Partial loss of wing margin results in proximodistal patterning defects (notches); complete loss results in loss of the wing blade²⁵.

Genetic analysis has identified a signalling pathway that regulates positioning of the wing margin. The putative secreted factor Fringe is expressed within the dorsal wing compartment²⁷. The juxtaposition of *fringe*-expressing and non-expressing cells seems to be necessary and sufficient for margin formation within the context of the wing disc. Significantly, outgrowths induced by ectopic

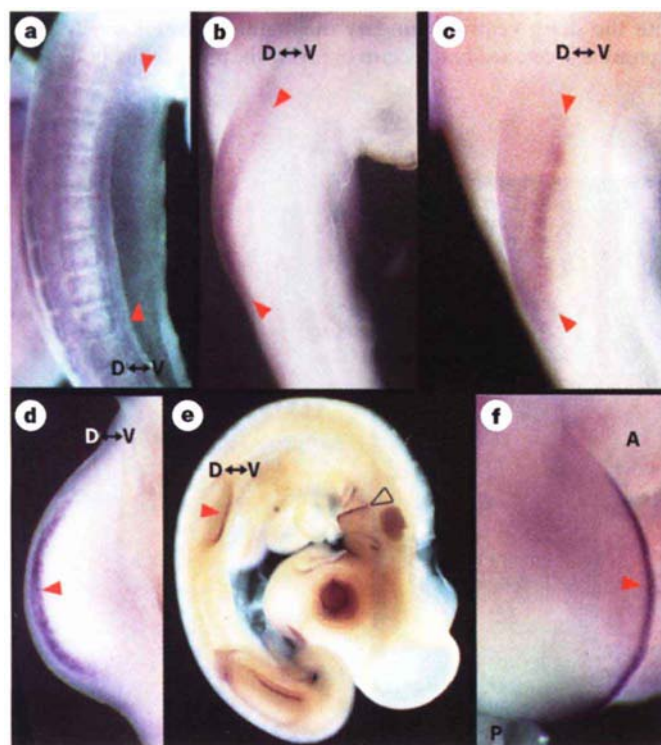


Figure 1 Expression patterns of *Radical fringe* and *Serrate-2* during early limb development. Whole-mount *in situ* hybridization showing *Radical fringe* (a–e) and *Serrate-2* (f) expression. *Radical fringe* expression is detected before AER formation in presumptive dorsal limb ectoderm by stage 15 (a). As limb outgrowth commences, this expression is maintained (b) and upregulated along the dorsal/ventral (D ↔ V) border (red arrowheads in a–c) at stage 17 (c). After morphological AER formation, expression is detected primarily within the AER (d–f; red arrowheads). Expression is also detected at the posterior edge of branchial arch ectoderm (open arrowhead in e). f, *Serrate-2* is expressed in the AER of a stage-22 wing bud (red arrowhead; A, anterior; P, posterior).

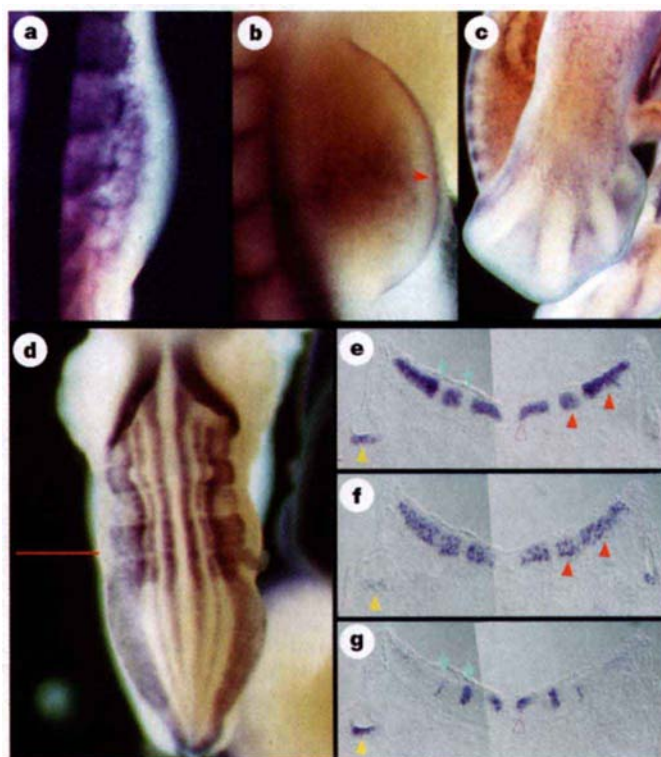
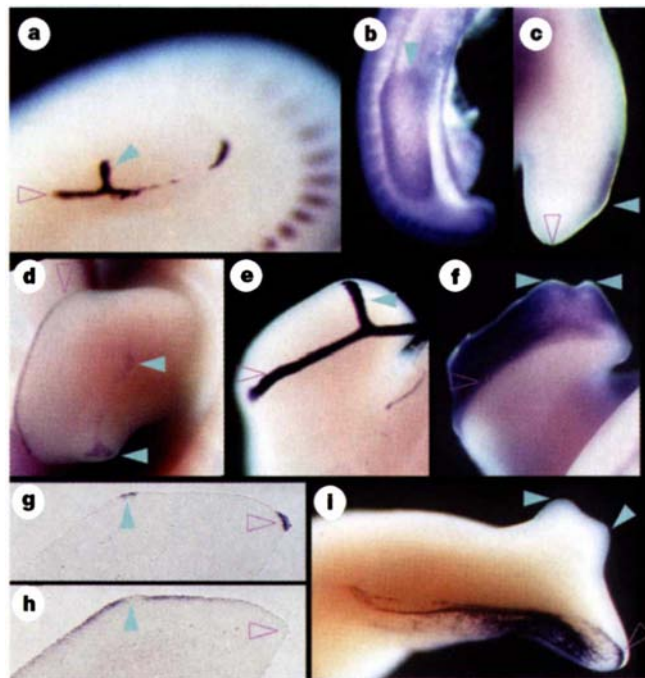


Figure 2 *Lunatic fringe* expression in the limb and central nervous system, and its relationship to Notch ligands. *Lunatic fringe* (a–c) expression is probably within angioblasts, as its distribution mirrors that of the developing limb vasculature from stage 18 (a) through stages 20 (b) and 29 (c). Expression in the sub-apical vessel is not likely to be involved in AER induction (red arrowhead in b), as this vessel forms later than the AER. d, e, *Lunatic fringe* is expressed in longitudinal stripes in the neural tube, shown here in whole mount of a stage-18 embryonic hindbrain. These stripes (e) correlate with the stripes of expression noted³⁰ for *Delta-like-1* (f) and *Serrate-1* (g), as revealed in serial transverse sections of a stage-23 embryonic hindbrain at the anteroposterior level of the otocyst (approximate plane of section indicated by red line in d). The lateral expression domains of *Lunatic fringe* and *Delta-like-1* overlap (filled red arrowheads) and are complementary with those of *Serrate-1* (filled blue arrows), while medially *Lunatic fringe* and *Serrate-1* expression overlap (open magenta arrowheads). All three genes are expressed in similar domains in the otocyst (filled yellow arrowheads).

Our analysis of AER formation indicates that positioning of the AER is indeed regulated similarly to that of the *Drosophila* wing margin. We have isolated two chick homologues of *fringe*: *Radical fringe* and *Lunatic fringe*. We have analysed expression patterns of the vertebrate *fringe* genes in both wild-type and AER-defective chicken strains, performed grafting experiments juxtaposing *fringe*-expressing and non-expressing limb ectoderm, and misexpressed the vertebrate *fringe* and *Engrailed-1* genes within developing chick limb buds. Our results support the hypothesis that the AER forms at the juxtaposition of *Radical fringe*-expressing and non-expressing cells, and furthermore dissociate the molecular control of AER formation from that of dorsoventral tissue specification. They also suggest that *Engrailed-1* is indirectly involved in AER formation through regulation of *Radical fringe* expression.

The analogy between AER and *Drosophila* wing margin function, and the parallel between their locations at their respective dorso-

To investigate whether either of these vertebrate *fringe* genes might be involved in AER formation, their expression patterns were examined by *in situ* hybridization. At stage 15/16, before limb-bud initiation, *Radical fringe* mRNA expression is detectable in the dorsal ectoderm along the primary body axis and in two broad lateral domains situated at axial levels corresponding to the presumptive limb fields (Fig. 1a; and data not shown). Within the limb field domains, a sharp demarcation of expression is seen, such that the more lateral regions of the ectoderm, which will give rise to ventral limb ectoderm, have greatly reduced levels of *Radical fringe* expression. As the limb buds emerge at stage 17 (Fig. 1b), high levels of *Radical fringe* expression remain confined to the dorsal ectoderm, with the sharp ventral boundary maintained. More laterally still, expression is detected in ectoderm overlying the presumptive body-wall



surface at stage 27 (**d**). *Fgf-8* expression marks ectopic dorsal and normal AERs of stage-25 mutant limb buds (**e**). Note that some ectopic AERs form perpendicular to the primary AER (**e**), while others run parallel to it (**d**). *Lmx-1* expression in mutant limb buds is detected throughout the dorsal mesoderm (**f**). *Wnt-7a* is expressed in the dorsal limb ectoderm of a mutant limb bud (**h**), on either side of an ectopic AER marked by *Fgf-8* expression (**g**). *En-1* is expressed in the ventral limb ectoderm and normal AER of a mutant limb bud, but not in association with the ectopic AERs (**i**). Normal AERs are indicated by open magenta arrowheads, ectopic AERs by filled green arrowheads. **a, e, f-i**, Dorsal top; **b, dorsal left; c, d**, dorsal right. **a-f, i**, Whole-mount embryos; **g, h**, transverse serial sections.

mesoderm (data not shown). As the limb bud continues to grow there are many changes in *Radical fringe* expression which presage overt AER formation. Expression becomes elevated on the dorsal side of the dorsoventral border, is downregulated throughout the rest of the dorsal ectoderm, and remains undetectable in the ventral limb ectoderm (Fig. 1c). Thus *Radical fringe* expression becomes confined primarily to the mature AER (Fig. 1d, e), where it is maintained until at least stage 27; low levels of dorsal ectodermal expression are detected until stage 24. *Radical fringe* expression is also detected in the somites, notochord and branchial arches (Fig. 1a, e; and data not shown).

Lunatic fringe has a complex and dynamic expression pattern throughout embryogenesis. Limb expression of *Lunatic fringe* is confined to a population of cells that appears to be migrating into the limb mesenchyme from stage 17 (Fig. 2a), and subsequently establish patterns suggestive of the limb vasculature (Fig. 2b, c). It is likely that these are angioblasts, as *Lunatic fringe* is expressed in a subset of vascular cells throughout the embryo (data not shown). Other sites of *Lunatic fringe* expression include the ventricular zone of the neural tube, where it is expressed in a series of longitudinal stripes (Fig. 2d, e) and the somites (data not shown). The *Lunatic fringe* expression pattern will be described in detail elsewhere.

Vertebrate fringe genes and Notch signalling

The formation of the AER adjacent to the dorsoventral border of *Radical fringe* expression is analogous to the formation of the wing margin at the dorsoventral border of fringe expression in the *Drosophila* wing imaginal disc. The inductive effects of Fringe in the *Drosophila* wing margin are mediated by downstream ligands in the Notch pathway, Delta and Serrate. To see whether *Radical fringe* might act through a similar pathway in the vertebrate limb, we examined the expression of several chicken homologues of Delta and Serrate. We find that one of these, *Serrate-2*, is expressed in the AER from the earliest stages of its formation through at least stage 26 (Fig. 1f; and data not shown). Chicken *Notch-1* is also expressed in the AER³⁰. Thus *Radical fringe*, like its *Drosophila* counterpart, may act upstream of Notch.

The expression of Notch ligands is also associated with that of *Lunatic fringe*. For example, *Delta-like-1* is expressed in longitudinal stripes in the neural tube that generally overlap those of *Lunatic fringe*, whereas *Serrate-1* is expressed in longitudinal stripes that are generally complementary to the *Lunatic fringe* stripes³⁰ (Fig. 2e–g). This relationship is not absolute, however, as in some regions of the neural tube, such as near the floorplate, the converse overlaps are observed. These three genes are also expressed in complex overlapping patterns within the otocyst (Fig. 2e–g) and other regions of the embryo, including the blood vessels and somites³⁰ (data not shown). Thus there is a general association of Notch signalling and vertebrate fringe gene expression.

Radical fringe expression patterns

The localization of *Radical fringe* expression to the dorsal ectoderm is particularly intriguing as the AER in normal limb buds is positioned at the juxtaposition of dorsal and ventral ectodermal tissue. To test whether the apposition of dorsal and ventral limb-bud ectoderm is sufficient to lead to AER formation, we grafted either ventral or dorsal ectoderm onto the ectoderm-denuded dorsal surface of stage 18/19 host leg buds. Of the ventral-to-dorsal grafts, 77% (23 of 29) induced ectopic AERs (Fig. 3a), but none of the dorsal-to-dorsal control grafts did. The importance of the dorsoventral border is also seen in *limbless* homozygote limb buds, which have previously been shown to be bidorsal in character and lack an AER. In these limb buds, *Radical fringe* expression is detected throughout the limb-bud ectoderm (Fig. 3b). These results are consistent with two possibilities: that the AER forms at the juxtaposition of dorsally and ventrally fated tissues, or that it is dictated by the *Radical fringe* expression boundary.

We were able to distinguish between these possibilities by examining gene expression in *eudiplopodia* mutant chick embryos. Ectopic AERs in *eudiplopodia* homozygote embryos are initially visible at stages 22–23, and produce very obvious dorsal outgrowths by stage 25. Ectopic patches of upregulated *Radical fringe* expression are detected on the dorsal surface of *eudiplopodia* homozygote limb buds from stage 22, and are larger than, and associated with, nascent outgrowths (Fig. 3c). By stages 25–26, *Radical fringe* expression is restricted to the ectopic and endogenous AERs (Fig. 3d). The ectopic AERs in *eudiplopodia* mutant embryos tend to form along one of two orthogonal axes. As judged by expression of the AER-specific marker Fgf-8, the ectopic AERs form primarily either parallel or perpendicular to the position of the normal AER (Fig. 3d, e).

The outgrowths that result from ectopic *eudiplopodia* AERs are morphologically bidorsal, suggesting that they may not arise from the juxtaposition of dorsal and ventral ectoderm. This was tested by examining the expression of genes known to specify dorsoventral cell fate, and which also serve as dorsal and ventral molecular markers. The dorsal ectoderm marker Wnt-7a is expressed throughout *eudiplopodia* homozygote dorsal ectoderm, but excluded from the ectopic AERs (Fig. 3g, h). The dorsal mesodermal marker Lmx-1 is similarly expressed throughout *eudiplopodia* homozygote dorsal mesoderm (Fig. 3f). In contrast, the ventral ectodermal marker En-1 is expressed only in *eudiplopodia* homozygote ventral ectoderm and the ventral half of the endogenous AER (Fig. 3i); it is not detectable in the ectopic AERs or the ectodermal surfaces of the ectopic outgrowths. Thus, in *eudiplopodia*, the specification of ectopic AER position occurs independently of the expression boundaries of genes known to impart dorsoventral cell fate, but correlates with *Radical fringe* expression boundaries.

Ectopic expression of vertebrate fringe

The expression pattern of *Radical fringe* in wild-type and *limbless* and *eudiplopodia* mutant limb buds suggests that the border between *Radical fringe*-expressing and non-expressing cells may be involved in specifying the position of the AER. If this is the case, abolishing this border by misexpression of fringe genes should prevent AER formation, and lead to a lack of outgrowth of the underlying mesoderm, whereas ectopic fringe expression domains

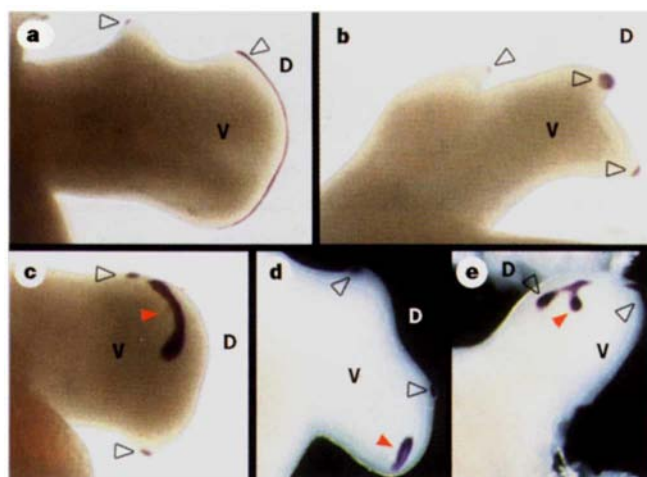


Figure 4 Fringe misexpression alters AER formation. Embryos infected with *Radical fringe* or *Lunatic fringe*-expressing replication-competent retroviral vectors at stages 9–11, collected from stages 23–27. **a–e**, Embryos stained by whole-mount *in situ* hybridization for Fgf-8 expression to mark the AER. Gaps in the normal AER (**a–e**), and ectopic ventral AER segments (filled red arrowheads) both parallel (**c, d**) and orthogonal (**e**) to the normal AER (open black arrowheads) were detected. D, dorsal; V, ventral.

might induce ectopic AERs. To test these predictions directly we misexpressed both *Lunatic fringe* and *Radical fringe* in developing limb buds using replication-competent retroviral vectors. Embryos were infected at stages 7–11 and allowed to develop until stages 22–28, after AER formation and considerable limb outgrowth had occurred. Viruses encoding either *fringe* gene produced limb buds with similar mutant phenotypes that were not observed with control retroviruses encoding alkaline phosphatase. Resultant phenotypes can be divided into three distinct classes, based on limb bud morphology and expression of the AER marker *Fgf-8*. The first class includes limb buds with disrupted AERs. Those segments of the distal ectoderm lacking a morphological AER also fail to express *Fgf-8*. The mesoderm underlying these regions fails to grow at the same rate as regions subjacent to an AER, resulting in limbs with a dramatically notched appearance (Fig. 4a, b). The second phenotypic class has gaps in the endogenous AER, as well as ectopic AER segments on the ventral surface supporting ectopic distal outgrowths (Fig. 4c, d). No ectopic AERs were observed on the dorsal surface of any infected limb bud. A subset of these ectopic AERs are perpendicular to the normal AERs, and adjacent to gaps within the normal AER, suggesting that a domain of misexpression simultaneously disrupted the endogenous *fringe* boundary and created a new one. These perpendicular AERs are remarkably similar to those seen on both the *eudiplopodia* mutant limbs and following ventral-to-dorsal ectoderm grafts (see Fig. 3). In the third class, embryos are missing an entire limb (data not shown).

To determine whether the alterations in AER patterning correlated with boundaries of *fringe* misexpression, several infected embryos with disrupted AERs were sectioned and stained for expression of *Fgf-8* as an AER marker and a retroviral protein epitope to delineate the domain of viral infection. In each case, boundaries of viral expression were found to correlate with the gain or loss of an AER. For example, we examined an embryo in which the anterior AER is disrupted, a small ectopic ventral AER was induced, and in the posterior the normal AER is intact. In the anterior, where the AER is lost, viral infection is observed throughout the dorsoventral extent of the anterior limb ectoderm, abolishing the *fringe* expression border (Fig. 5a, b). The first serial section

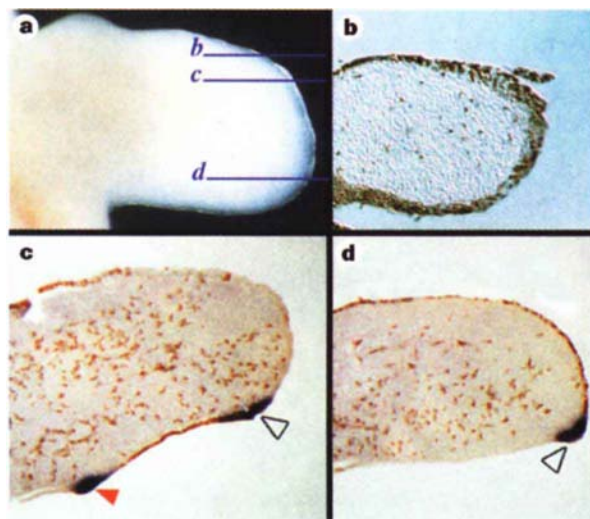


Figure 5 Fringe-induced changes in AER formation correlated with retroviral infection patterns. A limb bud infected with a *Radical fringe*-expressing retroviral vector sectioned and stained for expression of *Fgf-8* by non-radioactive *in situ* hybridization (blue stain; **c, d**) and immunohistochemically for retroviral p27^{gag} protein expression (brown stain; **b-d**). The approximate position of each section is indicated by the solid lines in **a**. **a**, Ventral view, anterior top; **b-d**, proximal left, dorsal top.

with a disrupted AER is also the first section showing viral expression on both the dorsal and ventral sides of the dorsoventral border. Where the ectopic ventral AER is present, it is found at the ventral edge of a patch of ventral ectodermal infection, which created a new border of *fringe*-expressing and non-expressing cells (Fig. 5c). More posteriorly, where the normal AER remains intact, the dorsal ectoderm is infected, in some regions including portions of the AER itself, but the ventral ectoderm is sparsely infected, if at all (Fig. 5d). In an embryo with only a single hind limb, heavy infection is detected throughout the affected side, suggesting that total limb-bud infection completely abolished the AER and hence abolished limb outgrowth (data not shown). These results are consistent with the hypothesis that the observed phenotypes result from ectopic ventral *fringe* expression altering the normal *Radical fringe* expression boundary.

Engrailed specifies AER position

En-1 is normally expressed in the ventral limb-bud ectoderm, and the abnormal AERs found in *en-1*^{-/-} mice indicate that it normally is involved in AER formation. We therefore investigated the effects

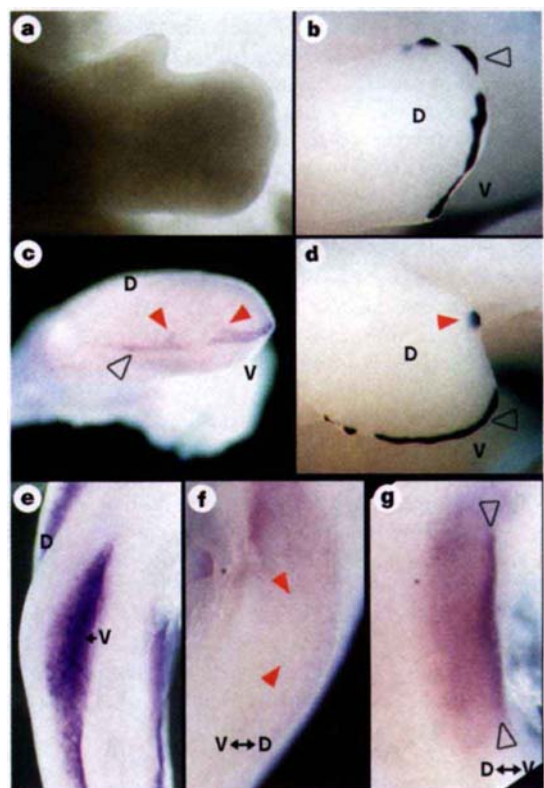


Figure 6 *Engrailed-1* regulates AER formation through repression of *Radical fringe* expression. **a-d**, RCAS-*En1* infected embryos allowed to develop to stages 23–26. Mesenchymal notching (**a**), gaps in the normal AER stained for *Fgf-8* (**b, d**) or *Radical fringe* (**c**) expression by whole-mount *in situ* hybridization, and ectopic AER segments on the dorsal limb surface, either continuous with (**c**) or independent of (**d**) the normal AER were observed. In addition, variations in the dorsoventral thickness of the normal AER were seen (**b**). Filled red arrowheads indicate ectopic AERs, and open black arrowheads indicate the normal AER, in **b-d**. **e**, Whole-mount *in situ* hybridization of a stage-17 embryo showing expression *En-1* in the ventral limb bud ectoderm. **f, g**, Infection with RCAS-*En1* at stage 10 followed by whole-mount *in situ* hybridization for *Radical fringe* expression reveals repression of *Radical fringe* expression. Filled red arrowheads in **f** indicate the expected position of the *Radical fringe* expression boundary, as compared to the normal expression boundary indicated by the open black arrowheads in **c**. D, dorsal; V, ventral.

Drosophila wing disc, *fringe* acts in part by inducing reciprocal signalling across the dorsoventral boundary by Serrate and Delta, ligands in the *Notch* signalling pathway²⁸. Vertebrate *serrate* and *Notch* homologues, *Ser-2* (Fig. 1f) and *Notch-1* (ref. 30) are expressed in the chick AER in a manner consistent with their being downstream of *Radical fringe* in a similar genetic cascade (Fig. 7), although this hypothesis requires experimental verification. The correspondence of vertebrate *fringe* expression with that of members of the *Notch* ligand family extends to other regions of the embryo, including the developing central nervous system, somites and blood vessels³⁰ (Fig. 2; and data not shown). These expression patterns suggest that *fringe* genes may generally act through modulation of the *Notch* signalling pathway.

Engrailed function

The AERs of *en-1*^{-/-} mice are elongated along the ventral limb ectoderm but have an appropriately positioned dorsal edge, suggesting that *En-1* is involved in positioning the ventral margin of the normal AER¹⁶. Our results, and those of C. Logan and A. Lumsden (personal communication), show that ectopic *En-1* can disrupt normal AER formation and give rise to dorsal AERs. We find that this probably occurs through repression of *Radical fringe* expression. However, a border of *En-1* expression is not necessary for AER formation, as ectopic dorsal AERs form in *eudiplopodia* homozygote mutant limbs independently of *En-1* expression. Taken together, these results suggest that *En-1* normally functions to position the ventral edge of the AER in the ventral ectoderm, possibly by regulating *Radical fringe* expression. Nevertheless, if a *Radical fringe* expression boundary is established by an *En-1*-independent mechanism, then an otherwise normal AER can develop.

AER forms independently of dorsoventral axis

It has been proposed³⁴ that AERs are induced by interactions occurring at the boundaries of the dorsal and ventral subdivision of vertebrate limb fields. However, the ectopic AERs present in *eudiplopodia* mutant embryos clearly establish that AER formation can occur independently of a dorsoventral fate boundary. Thus the dorsal markers *Wnt-7a* and *Lmx-1* are expressed throughout *eudiplopodia* dorsal ectoderm where ectopic AERs form, but the ventral *En-1* marker is excluded. Similar results with these markers have been obtained by B. Morgan (personal communication).

The normal AER forms at the boundary between dorsal and ventral limb ectoderm, albeit through the expression of *Radical fringe*. *En-1* apparently plays a key role in linking AER positioning and dorsoventral patterning, as it can repress expression of both *Radical fringe* and *Wnt-7a* (C. Logan and A. Lumsden, personal communication, and ref. 16), thereby limiting them to the dorsal ectoderm. The same genetic pathway thus influences both dorsoventral patterning and AER positioning, although it subsequently separates into two distinct branches (Fig. 7).

Induction of the AER

Grafting experiments indicate that AERs are induced by a signal from the limb-bud mesoderm. In recombination experiments between limb mesoderm and flank ectoderm, the ectoderm has been found to respond by forming an AER until about stage 20 (ref. 8). This competence period correlates well with the persistence of *Radical fringe* expression within the limb ectoderm. Because minced, recombined mesoderm works as effectively in AER induction as intact mesoderm⁵, this appears to be a permissive, rather than an instructive, signal. An AER is apparently induced by the limb mesoderm as long as there also exists a *Radical fringe* expression boundary along which the AER can form. This inductive signalling pathway probably involves a member of the FGF family, as FGF-soaked beads implanted into stage-18 interlimb flank induce limb outgrowth and AER formation.

Evolutionary implications

The role of *Radical fringe* in the chick limb is remarkably similar to the function of *Drosophila fringe* within the wing disc. In each case *fringe* genes are acting to set a boundary within a two-dimensional epithelium to locate a linear signalling centre that controls proximodistal outgrowth. This suggests that the wing margin and the AER are functionally homologous structures formed by the action of conserved genetic pathways.

This is the second example of conserved signalling systems acting similarly during the development of both fly wings and amniote limbs. The anteroposterior axis of each is patterned through the action of the conserved *hedgehog/Sonic hedgehog* genetic pathways³⁵. The use of both the *fringe* and *hedgehog* pathways in patterning the limbs of members of these divergent phyla might be serendipitous. However, these two apparently independent genetic pathways might already have been used together to regulate the outgrowth and patterning of secondary fields in the last common ancestor of vertebrates and invertebrates. Both sets of signals are also present in the developing branchial arches³⁶, and the arch epithelium has AER-like activity³⁷, suggesting that their joint use in chordates at least pre-dates the evolution of paired appendages. □

Methods

Chick embryos. Fertile White Leghorn chicken eggs were purchased from SPAFAS, Inc. (Norwich, CT) or were from a flock maintained at the University of Wisconsin-Madison (for tissue grafting experiments). *Eudiplopodia* mutant embryos were produced by crossing heterozygous Single-comb White Leghorn chickens. Mutant embryos were identified by ectopic dorsal limb outgrowths visible from stage 22. *Limbleless* mutant embryos were obtained from a heterozygous mating flock maintained at the University of Wisconsin-Madison Poultry Science Department²². *Limbleless* mutant embryos were identified by lack of AERs at stage 18/19. All embryos were incubated at 38 °C in a humidified atmosphere.

Tissue grafts. Stage 18/19 donor embryos were lightly stained *in ovo* with neutral red to mark cells of the AER, limb buds were amputated, and ectodermal jackets isolated as described³⁸. Ventral ectodermal tissue grafts were dissected free of the ectodermal hull in 1 × EBSS + 10% horse serum, carefully excluding apical ridge cells retaining neutral red from the graft tissue. Stage 18/19 host leg bud dorsal surface graft beds were prepared, and the grafting technique performed as described³⁹.

Cloning of *fringe* and *serrate* genes. The human *fringe* EST clones ylc-3md09 and ylc-32h10 were identified by a TBLASTN screen⁴⁰ of the *Drosophila* fringe protein sequence against the dbEST database, and were obtained from the Genexpress cDNA Program (Laboratoire Genethon, Evry, France). Plaques (10⁶) of a s12-15 embryonic chick cDNA library cloned in λZAPII were screened, and clones were recovered in pBluescript SK+ (Stratagene, La Jolla, CA). Clones cFR30 and cFR42 were ultimately identified that contained the entire open reading frames of *Lunatic fringe* and *Radical fringe*, respectively, and were used for retroviral construction. Chicken *Ser-2* was cloned by RT-PCR³⁰. Computer analysis of DNA and protein sequences was performed using the DNASTAR software suite (Madison, WI). Standard recombinant DNA methodologies⁴¹ were used, and enzymes were purchased from Boehringer Mannheim Biochemicals (Indianapolis, IN).

In situ hybridization. Non-radioactive whole-mount and section *in situ* hybridizations of chick embryos were performed as described^{42,43}, taking care to titrate the proteinase K treatments to reveal ectodermal expression patterns. Clones cFR8 (*Lunatic fringe*, 2.1 kb insert, *Clal* digest, T3 polymerase) and cFR10 (*Radical fringe*, 3.4 kb, *Clal*, T3) were used as templates for *in situ* hybridization probes. The *Ser-2* probe is a 432-bp RT-PCR product. The probes for *Lmx-1* (ref. 13), *Wnt-7a* (ref. 44), *Fgf-8* (ref. 4), *Engrailed-1* (ref. 21), *Ser-1* (ref. 30) and *Delta-like-1* (ref. 30) have been described.

RCAS retroviruses. *Radical fringe* and *Lunatic fringe* cDNAs encoding the entire open reading frame of each gene with no 5' or 3' untranslated sequences were fused in frame with the initiator ATG of the pSLAX-13 shuttle vector. Influenza haemagglutinin (HA) or Flag (Kodak, Rochester, NY) epitope tags were then added to the C terminus of each protein. The tagged cDNAs were then transferred as *Clal* fragments into the RCASBP(B) retroviral vector. The

RCASBP(A) *En-1* virus has been described³¹. Retrovirus was prepared, concentrated and titted essentially as described⁴⁵. Retroviral titres ranged from 5×10^8 to 2×10^9 CFU ml⁻¹. Chick embryo fibroblasts infected with the tagged viruses were stained with antibodies to both retroviral p27^{gag} (3C2)⁴⁶ and HA (12CA5)⁴⁷ or Flag, and detected by indirect immunohistochemical methods. In all cases, viral infection and fringe protein expression were found to be >99% coincident. Stage 7–11 embryos were injected with concentrated viral supernatants, supplemented with 80 ng ml⁻¹ polybrene for B-coat viruses, along the flank and in the neural tube, allowing significant amounts to spill on the lateral plate ectoderm. Embryos were allowed to develop to stages 17–28, then fixed for further analysis.

Received 24 December 1996; accepted 6 March 1997.

1. Tickle, C. Vertebrate limb development. *Curr. Opin. Genet. Dev.* **5**, 478–484 (1995).
2. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Exp. Morphol.* **88**, 49–92 (1951).
3. Crossley, P. H., Minowada, G., MacArthur, C. A. & Martin, G. R. Roles for FGF8 in the induction, initiation, and maintenance of chick limb development. *Cell* **84**, 127–136 (1996).
4. Vogel, A., Rodriguez, C. & Izpisua-Belmonte, J. C. Involvement of FGF-8 in initiation, outgrowth and patterning of the vertebrate limb. *Development* **122**, 1737–1750 (1996).
5. Kieny, M. Role inducteur du mésoderme dans la différenciation précoce du bourgeon de membre chez l'embryon de poulet. *J. Embryol. Exp. Morphol.* **8**, 457–467 (1960).
6. Kieny, M. Variation de la capacité inductrice de mésoderme et de la compétence de l'ectoderme au cours de l'induction primaire du bourgeon de membre chez l'embryon de Poulet. *Arch. d'Anat. Microsc.* **57**, 401–418 (1968).
7. Saunders, J. Jr & Reuss, C. Inductive and axial properties of prospective wing-bud mesoderm in the chick embryo. *Dev. Biol.* **38**, 41–50 (1974).
8. Carrington, J. L. & Fallon, J. F. The stages of flank ectoderm capable of responding to ridge induction in the chick embryo. *J. Embryol. Exp. Morphol.* **84**, 19–34 (1984).
9. Goetinck, P. Studies on limb morphogenesis II. Experiments with the polydactylous mutant eudiplopodia. *Dev. Biol.* **10**, 71–91 (1964).
10. Fraser, R. A. & Abbott, U. K. Studies on limb morphogenesis. VI. Experiments with early stages of the polydactylous mutant Eudiplopodia. *J. Exp. Zool.* **176**, 237–248 (1971).
11. Parr, B. A., Shea, M. J., Vassileva, G. & McMahon, A. P. Mouse Wnt genes exhibit discrete domains of expression in the early embryonic CNS and limb buds. *Development* **119**, 247–261 (1993).
12. Vogel, A., Rodriguez, C., Warnken, W. & Izpisua Belmonte, J. C. Dorsal cell fate specified by chick *Lmx1* during vertebrate limb development. *Nature* **378**, 716–720 (1995).
13. Riddle, R. D. et al. Induction of the LIM homeobox gene *Lmx-1* by *Wnt-7a* establishes dorsoventral pattern in the vertebrate limb. *Cell* **83**, 631–640 (1995).
14. Parr, B. A. & McMahon, A. P. Dorsalizing signal *Wnt-7a* required for normal polarity of D–V and A–P axes of mouse limb. *Nature* **374**, 350–353 (1995).
15. Davis, C. A., Holmyard, D. P., Millen, K. J. & Joyner, A. L. Examining pattern formation in mouse, chicken and frog embryos with an En-specific antiserum. *Development* **111**, 287–298 (1991).
16. Loomis, C. A. et al. The mouse *Engrailed-1* gene and ventral limb patterning. *Nature* **382**, 360–363 (1996).
17. Prahlad, K. V., Skala, G., Jones, D. G. & Briles, W. E. Limbless: a new genetic mutant in the chick. *J. Exp. Zool.* **209**, 427–434 (1979).
18. Fallon, J. F., Frederick, J. M., Carrington, J. L., Lanser, M. E. & Simandl, B. K. Studies on a limbless mutant in the chick embryo. *Prog. Clin. Biol. Res.* **110**, 33–43 (1983).
19. Rosenblatt, L., Kreutziger, G. & Taylor, L. Eudiplopodia. *Poultry Sci.* **38**, 1242 (1959).
20. Carrington, J. L. & Fallon, J. F. Initial limb budding is independent of apical ectodermal ridge activity; evidence from a limbless mutant. *Development* **104**, 361–367 (1988).
21. Noramly, S., Pisenti, J., Abbott, U. & Morgan, B. Gene expression in the limbless mutant: Polarized gene expression in the absence of Shh and an AER. *Dev. Biol.* **179**, 339–346 (1996).
22. Ros, M. A. et al. The limb field mesoderm determines initial limb bud anteroposterior asymmetry and budding independent of sonic hedgehog or apical ectodermal gene expressions. *Development* **122**, 2319–2330 (1996).
23. Grieshammer, U., Minowada, G., Pisenti, J. M., Abbott, U. K. & Martin, G. R. The chick limbless mutation causes abnormalities in limb bud dorsal–ventral patterning: implications for the mechanism of apical ridge formation. *Development* **122**, 3851–3861 (1997).
24. Tanaka, M., Tamura, K., Noji, S., Nohno, T. & Ide, H. Induction of an additional limb at the dorsal–ventral boundary of a chick embryo. *Dev. Biol.* **182**, 191–203 (1997).
25. Diaz-Benjumea, F. J. & Cohen, S. M. Interaction between dorsal and ventral cells in the imaginal disc directs wing development in *Drosophila*. *Cell* **75**, 741–752 (1993).
26. Blair, S. S. Mechanisms of compartment formation: evidence that non-proliferating cells do not play a critical role in defining the D/V lineage restriction in the developing wing of *Drosophila*. *Development* **119**, 339–351 (1993).
27. Irvine, K. D. & Wieschaus, E. fringe, a Boundary-specific signaling molecule, mediates interactions between dorsal and ventral cells during *Drosophila* wing development. *Cell* **79**, 595–606 (1994).
28. Brook, W. J., Diaz-Benjumea, F. J. & Cohen, S. M. Organizing spatial pattern in limb development. *Annu. Rev. Cell. Dev. Biol.* **12**, 161–180 (1996).
29. Wu, J. Y., Wen, L., Zhang, W. J. & Rao, Y. The secreted product of *Xenopus* gene *lunatic fringe*, a vertebrate signaling molecule. *Science* **273**, 355–358 (1996).
30. Myatt, A., Henrique, D., Ish-Horowitz, D. & Lewis, J. A chick homologue of Serrate and its relationship with Notch and Delta homologues during central neurogenesis. *Dev. Biol.* **174**, 233–247 (1996).
31. Logan, C. et al. Rostral optic tectum acquires caudal characteristics following ectopic Engrailed expression. *Curr. Biol.* **6**, 1006–1014 (1996).
32. Rodriguez, C. & Izpisua-Belmonte, J. C. (submitted).
33. Raser, R. A. & Abbott, U. K. Studies on limb morphogenesis. V. The expression of Eudiplopodia and its experimental modification. *J. Exp. Zool.* **176**, 219–236 (1971).
34. Meinhardt, H. A boundary model for pattern formation in vertebrate limbs. *J. Embryol. Exp. Morphol.* **76**, 115–137 (1983).
35. Ingham, P. W. Pattern formation. Hedgehog points the way. *Curr. Biol.* **4**, 347–350 (1994).
36. Wall, N. A. & Hogan, B. L. Expression of bone morphogenetic protein-4 (BMP-4), bone morphogenetic protein-7 (BMP-7), fibroblast growth factor-8 (FGF-8) and sonic hedgehog (SHH) during branchial arch development in the chick. *Mech. Dev.* **53**, 383–392 (1995).
37. Richman, J. M. & Tickle, C. Epithelial-mesenchymal interactions in the outgrowth of limb buds and facial primordia in chick embryos. *Dev. Biol.* **154**, 299–308 (1992).
38. Ros, M. A., Lyons, G. E., Mackem, S. & Fallon, J. F. Recombinant limbs as a model to study homeobox gene regulation during limb development. *Dev. Biol.* **166**, 59–72 (1994).
39. Carrington, J. L. & Fallon, J. F. Experimental manipulation leading to induction of dorsal ectodermal ridges on normal limb buds results in a phenocopy of the Eudiplopodia chick mutant. *Dev. Biol.* **116**, 130–137 (1986).
40. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
41. Ausubel, F. M. et al. *Current Protocols in Molecular Biology* (Greene/Wiley Interscience, New York, 1989).
42. Riddle, R. D., Johnson, R. L., Laufer, E. & Tabin, C. Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* **75**, 1401–1416 (1993).
43. Yang, X. & Cepko, C. L. Flk-1, a receptor for vascular endothelial growth factor (VEGF), is expressed by retinal progenitor cells. *J. Neurosci.* **16**, 6089–6099 (1996).
44. Dealy, C. N., Roth, A., Ferrari, D., Brown, A. M. & Koshier, R. A. Wnt-7a are expressed in the developing chick limb bud in a manner suggesting roles in pattern formation along the proximodistal and dorsoventral axes. *Mech. Dev.* **43**, 175–186 (1993).
45. Morgan, B. A. & Fekete, D. M. Manipulating gene expression with replication-competent retroviruses. *Methods Cell. Biol.* **51**, 185–218 (1996).
46. Potts, W. M., Olsen, M., Boettiger, D. & Vogt, V. M. Epitope mapping of monoclonal antibodies to gag protein p19 of avian sarcoma and leukaemia viruses. *J. Gen. Virol.* **68**, 3177–3182 (1987).
47. Wilson, I. A. et al. The structure of an antigenic determinant in a protein. *Cell* **37**, 767–778 (1984).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.wm>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank B. K. Simandl for technical help and reading the manuscript. The *En-1* virus was a gift from C. Logan and A. Lumsden; the *Wnt-7a* probe was from A. Roth and T. Brown; the *Fgf-8* was from J. C. I. Belmonte; the cDNA library was from A. Nieto and D. Wilkinson; and we thank J. C. I. Belmonte, B. Morgan, C. Logan and A. Lumsden for sharing results before publication. This work was supported by an NSF predoctoral fellowship to R.D., and funding from the NIH to J.F.F. and the American Cancer Society to C.T.

Correspondence and requests for materials should be addressed to C.T. (e-mail: tabin@rascal.med.harvard.edu). The Genbank accession nos are as follows: *Lunatic fringe*, U91849; *Radical fringe*, U91850; *Serrate-2*, U91851.

Anomalous small magnetic field in the local interstellar cloud

George Gloeckler^{*†}, L. A. Fisk[†] & Johannes Geiss[‡]

^{*} Department of Physics, University of Maryland, College Park, Maryland 20742, USA

[†] Department of Atmospheric, Oceanic, and Space Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA

[‡] International Space Science Institute, Hallerstrasse 6, CH-3012 Bern, Switzerland

The solar wind carves out a cavity, known as the heliosphere, in the warm local interstellar cloud, which is itself embedded in a larger hot cloud. It is generally assumed that there is an overall pressure balance between these three regions. Thermal pressure and magnetic field pressure in the local interstellar cloud should therefore balance the inward pressure from the hot cloud^{1–3}, and determine the size of the heliosphere. Here we present direct measurements of the density and speed of interstellar hydrogen and helium ions deep inside the Solar System, from which we derive the thermal pressure in the local interstellar cloud. Combined with the fact that the magnetic field strength in the local cloud is constrained to be less than 4.3 μG (to be compatible with the fact that the Voyager 1 spacecraft has yet to encounter the heliosphere termination shock), the total pressure that we infer is insufficient to balance the inward pressure from the hot cloud. We conclude that either the magnetic field in the local cloud is inhomogeneous, or there is a significant, as yet undetected, non-thermal component to the pressure in the local cloud.

Interstellar hydrogen and helium atoms enter the Solar System owing to its relative motion with respect to the local interstellar cloud. Influenced only by solar gravity, these neutral species penetrate deep into the heliosphere, where some are ionized by solar radiation. We detect these ions (called pickup ions^{4–7}) with the Ulysses spacecraft in the inner heliosphere between 2.5 and 5 AU from the Sun, and use these measurements to obtain densities of interstellar neutral and ionized hydrogen and helium in the very local interstellar cloud, that is, within a few thousand AU from the Sun.

The density of interstellar neutral hydrogen reaching the inner heliosphere is significantly reduced by ionization in two distinct locations. The first, called the filtration region, is in the local interstellar cloud (LIC) just beyond the heliopause, a boundary separating solar from interstellar plasma. In this region (roughly 100–200 AU wide^{8–11}) the LIC plasma, forced to move around the heliopause, slows down and in a process known as filtration^{8–10}, converts some fraction of the H atoms to fast protons by proton–hydrogen charge-exchange. The ram and thermal pressure of the newly created protons and the original LIC plasma, combined with the pressure of the LIC magnetic field draped around the heliopause, determine the size of the heliosphere. The amount of reduction of atomic hydrogen passing through the filtration region is roughly proportional to its average proton density.

The second location where ionization occurs is inside the heliosphere, where both interstellar atomic H and He are ionized by the solar wind and solar ultraviolet radiation; the ionization rate (which we measure directly) is strongest near the Sun. The newly ionized particles, called pickup ions, are accelerated and swept away from the Sun by the solar wind. It is possible to observe pickup He^{2+} and H^+ and distinguish them from the far more abundant solar-wind He^{2+} and H^+ by their speed distribution as shown in Figs 1 and 2.

Figure 1 shows the speed distribution of He^{2+} measured with the Solar Wind Ion Composition Spectrometer¹² (SWICS) on the Ulysses spacecraft. Pickup He^{2+} is produced predominantly by double charge-exchange with solar-wind α -particles and its rate of production from neutral He is known from SWICS measurements of the solar-wind He^{2+} flux. The best fit to the observed pickup He^{2+} spectrum both above and below the solar-wind peak yields an interstellar neutral helium density, $n(\text{He}_{\text{TS}}) = 0.0153 \pm 0.0018 \text{ cm}^{-3}$. Our result is in excellent agreement with the value of $0.0155 \pm 0.0015 \text{ cm}^{-3}$ derived from direct measurements of interstellar neutral helium on Ulysses¹³. Because He is not affected by filtration, $n(\text{He}_{\text{I}})$, the neutral-helium density in the LIC, is equal to $n(\text{He}_{\text{TS}})$.

The distribution function of protons measured near the ecliptic plane at distances ~ 4.5 AU from the Sun is shown in Fig. 2. Using the 'hot' model of Thomas¹⁴ with an interstellar gas temperature of 7,500 K, we calculate the expected distribution of pickup H^+ (solid curve). The shape of the distribution fixes the only free parameter of the model^{6,7}, β_{Hloss} , the rate at which hydrogen atoms are ionized at 1 AU in the heliosphere. The phase space density between 1.5 and twice the solar wind speed ($1.5 < W < 2$) gives the product of the density of interstellar neutral hydrogen in the outer heliosphere, $n(\text{H}_{\text{TS}})$, and the rate of production of pickup H^+ . From the best fit to the observed pickup H^+ distribution we find $n(\text{H}_{\text{TS}}) = 0.115 \pm 0.025 \text{ cm}^{-3}$. Previously, $n(\text{H}_{\text{TS}})$ has been deduced from observations of resonant back-scattering of solar ultraviolet radiation^{3,15}, but these methods depend on complex back-scattering models with poorly known parameters, and often have systematic errors due to calibration uncertainties¹⁶. Neutral-hydrogen densities based on ultraviolet observations, revised for more realistic β_{Hloss} rates¹⁵, are widely scattered from 0.03 to 0.3 cm^{-3} with an average value of $\sim 0.135 \text{ cm}^{-3}$, which agrees with our result.

Combining our values for the densities of atomic H and He at the termination shock with measured parameters in the LIC listed in Table 1, we evaluate the atomic H density in the LIC, $n(\text{H}_{\text{I}})$, and the average proton density in the filtration region, $\langle n(\text{p}) \rangle$. The computational steps and results are given in Table 1. We find that $n(\text{H}_{\text{I}}) = 0.20 \pm 0.03 \text{ cm}^{-3}$, and that $\sim 40\%$ of the LIC atomic hydrogen is excluded from the heliosphere by charge exchange in the filtration region.

A LIC magnetic field of $B < 3.7 \mu\text{G}$, when taken with values (T , R_2 and $n(\text{H}_{\text{II}}) = \langle n(\text{p}) \rangle$) from Table 1, shows that the Sun moves supersonically relative to the LIC. Assuming $B < 3.7 \mu\text{G}$, there will be a bow shock at the outer boundary of the filtration region. The magnetic field orientation observed to be nearly parallel¹⁷ to the shock surface will make it a perpendicular MHD shock. Using standard jump conditions¹⁸, and parameters from Table 1, we find the total LIC pressure, $P_{\text{LIC}}(r)$, at the heliopause as a function of the shock compression ratio, r ; we use the magnetic field and densities of atomic hydrogen, $n(\text{H}_{\text{I}})$, protons, $n(\text{H}_{\text{II}})$, and ionized helium, $n(\text{He}_{\text{II}})$ in the LIC just beyond the bow shock. The location, R_{TS} , of the termination shock (a boundary where the supersonic solar wind becomes subsonic) is found from pressure balance between the solar wind and the LIC plasmas at the heliopause. The solar-wind pressure at the heliopause, $P_{\text{SW}}(R_{\text{TS}})$, depends on the solar-wind density and speed just inside the termination shock and decreases rapidly with R_{TS} . Thus all LIC parameters, such as B , calculated as a function of the compression ratio may also be expressed in terms of R_{TS} . (See Fig. 3 legend for computational details.)

In Fig. 3a we show the dependence of the computed magnetic field on the location of the heliospheric termination shock. The shaded vertical region indicates the range of R_{TS} values derived from observations¹⁹ of anomalous cosmic rays with Voyagers 1 and 2; the minimum distance to the termination shock is at 80 AU. This places an upper limit on B of 3 μG (upper dashed curve in Fig. 3a), justifying our assumption of supersonic flow. Our most probable calculated field (solid curve) agrees remarkably well with reported

Table 1 Measured and derived average parameters in the local interstellar cloud

Parameter	Units	Local* interstellar cloud	Very local interstellar cloud ($\leq 2,000$ AU)
$n(\text{H}_{\text{TS}})$	cm^{-3}		$0.115 \pm 0.025^{\dagger}$
$n(\text{He}_{\text{I}}) = n(\text{He}_{\text{TS}})$	cm^{-3}		$0.0153 \pm 0.0018^{\dagger}$
V_0	km s^{-1}		$25.3 \pm 0.4^{\ddagger}$
T	K	$7,000 \pm 600^{\ddagger}$	
$R1 = n(\text{H})/n(\text{He}_{\text{I}})$		$13 \pm 1^{\S}$	
$R2 = n(\text{H})/n(\text{He})$		10	
$n(\text{H}_{\text{I}}) = R1 \times n(\text{He}_{\text{I}})$	cm^{-3}		0.20 ± 0.03
$F = n(\text{H}_{\text{TS}})/n(\text{H}_{\text{I}}) = \text{filtration factor}$			0.58 ± 0.15
$\langle n(p) \rangle = (-1/L\sigma)\ln(F/1.46)$	cm^{-3}		$0.073 \pm 0.031^{\P}$
$n(\text{H}_{\text{I}}) = n(p)/r$	cm^{-3}		$0.040 \pm 0.017^{\#}$
$n(\text{H}) = n(\text{H}_{\text{I}}) + n(\text{H}_{\text{II}})$	cm^{-3}		$0.240 \pm 0.034^{\#}$
$n(\text{He}_{\text{I}}) = [n(\text{H})]/R2 - n(\text{He}_{\text{II}})$	cm^{-3}		$0.009 \pm 0.004^{\#}$
$n_{\text{e}} = n(\text{H}_{\text{II}}) + n(\text{He}_{\text{II}})$	cm^{-3}		$0.049 \pm 0.016^{\#}$
$X_{\text{H}} = n(\text{H}_{\text{I}})/n(\text{H})$		$0.17^{\star} < X_{\text{H}} < 0.4^{\S}$	$0.18 \pm 0.01^{\star\star}$
$X_{\text{He}} = n(\text{He}_{\text{I}})/R2/n(\text{H})$		$0.3^{\star} < X_{\text{He}} < 0.5^{\S}$	$0.36 \pm 0.02^{\star\star}$

* Average values measured or calculated over distances of a few to ~ 100 pc and assumed to be also valid in the immediate neighbourhood of the heliosphere.

[†] Densities near the termination shock, present work.

[‡] From refs 13, 24.

[§] From ref. 21.

|| From ref. 25.

[¶] Proton density in the filtration region: the width of the filtration region is $L = 160 \pm 40$ AU (refs 8–11), and $\sigma = 5.3 \times 10^{-15} \text{ cm}^2$ is the proton–hydrogen charge-exchange cross-section²⁶ at $V_{\text{rel}} = 19 \text{ km s}^{-1}$, the average relative speed between protons and H atoms in the filtration region. The factor of 1.46 = $V_{\text{rel}}/(V_{\text{rel}} - 6)$ accounts for that fraction of slow-moving H atoms, produced by charge exchange, that enter the heliosphere, contribute to $n(\text{H}_{\text{TS}})$, and reduce the average speed of interstellar H inside the heliosphere²⁷ to 20 km s^{-1} .

[#] For compression ratio, $r = 1.8$.

[☆] From ref. 22.

^{☆☆} For $80 < R_{\text{TS}} < 85$ AU see Fig. 3 legend for error limits.

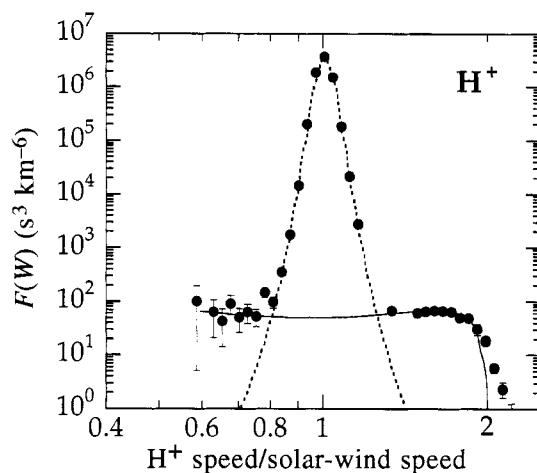
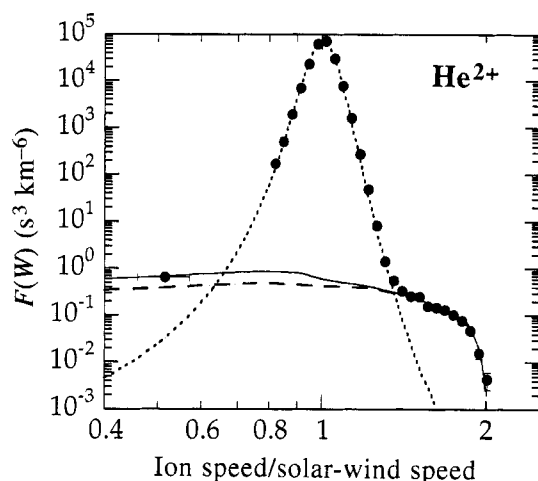


Figure 1 Phase space density of He^{2+} versus W , the ion speed in the spacecraft frame of reference divided by the solar-wind speed. This time-averaged spectrum was observed with SWICS for 111 days (19 July to 8 October, 1994) at ~ 2.5 AU and $\sim 70^\circ$ latitude in the persistently quiet high-speed ($\sim 808 \text{ km s}^{-1}$) solar wind. Long observation periods and quiet solar-wind conditions were required to observe these rare pickup ions. The dotted curve is the fit to the solar-wind He^{2+} , yielding a density of 0.022 cm^{-3} and a thermal speed of 29 km s^{-1} . The dashed curve is computed assuming that the pickup ion distribution function is isotropic. For the heliospheric spatial distribution of interstellar atomic He we use the ‘hot’ model¹⁴ with μ (radiation pressure divided by the gravitation force) of 0, a loss rate⁷ of $0.6 \times 10^{-7} \text{ s}^{-1}$, and He with speed of 26 km s^{-1} and temperature of $7,500 \text{ K}$. The production rate is measured to be $2.0 \times 10^{-9} \text{ s}^{-1}$, being almost entirely due to double charge-exchange with solar-wind α -particles ($\sigma = 1.8 \times 10^{-16} \text{ cm}^2$)²⁸. Because at high latitudes the average magnetic field is nearly radial, anisotropic pickup ion distributions are expected^{29,30}. The best fit (solid curve) for $1.7 < W < 2$ is obtained with $n(\text{He}_{\text{TS}}) = 0.0153 \pm 0.018 \text{ cm}^{-3}$ using a 25% anisotropy correction. We did not use here the far more abundant He^+ , discovered by Möbius⁴ and also observed with SWICS, to find $n(\text{He}_{\text{TS}})$ because during the time period of our observations its production rate from photoionization was not known owing to lack of direct measurements. Details of how the observed and model distributions are computed are given in refs 6 and 7.

Figure 2 As Fig. 1 for H^+ . This spectrum, $F(W)$, was measured during a 16-day quiet period in late 1991 when Ulysses was at ~ 4.88 AU in the ecliptic plane. In this period of relatively cool solar wind, the pickup H^+ is well separated from the far more abundant solar-wind protons centred at $W = 1$. The dotted curve is the fit to the solar-wind distribution using a convected maxwellian distribution with density of 0.09 cm^{-3} and thermal speed of 10.5 km s^{-1} . As the average heliospheric magnetic field direction was nearly perpendicular to the solar-wind flow direction, the pickup ion distribution was essentially isotropic, and no anisotropy corrections^{29,30} were required in determining interstellar densities. The solid curve is the fit to the pickup ion spectrum. We use the ‘hot’ model¹⁴ to compute the heliospheric spatial distribution of neutral interstellar hydrogen with a speed²⁷ of 20 km s^{-1} and temperature of $7,500 \text{ K}$; μ is taken as 1, and the isotropic form of the pickup ion speed distribution in the solar-wind frame³⁰ is used. With $\mu = 1$ the shape of the observed $F(W)$ for $1.25 < W < 1.9$ restricts β_{loss} , the loss rate of atomic H, to $(8.5 \pm 1) \times 10^{-7} \text{ s}^{-1}$. The best fit is obtained with $n(\text{H}_{\text{TS}}) = 0.115 \pm 0.025 \text{ cm}^{-3}$ using $\beta_{\text{up,prod}} = 2.7 \times 10^{-7} \text{ s}^{-1}$. The charge-exchange portion of $\beta_{\text{H,prod}}$ equals $1.7 \times 10^{-7} \text{ s}^{-1}$, and is the product of the observed solar-wind flux (adjusted to 1 AU) and the proton–hydrogen charge-exchange cross-section²⁶, $\sigma_p = 1.8 \times 10^{-15} \text{ cm}^2$ at the solar-wind speed of 475 km s^{-1} . Photoionization³¹ accounts for the rest.

values^{17,20} of the interstellar magnetic field of $1.4 \pm 0.15 \mu\text{G}$ represented by the shaded horizontal region.

The variation with R_{TS} of the fractional ionization of hydrogen, $X(\text{H})$, and helium, $X(\text{He})$, and the electron density, n_e , in the very local interstellar cloud is shown in Fig. 3b. Our values for both $X(\text{H})$ and $X(\text{He})$ are well below measured²¹ upper limits given in Table 1. However, imposing the calculated lower limit²² of 0.17 on $X(\text{H})$ places an upper limit on R_{TS} of $\sim 85 \text{ AU}$. Thus, the most probable value of B is between ~ 1.3 and $\sim 2 \mu\text{G}$ for the case of supersonic

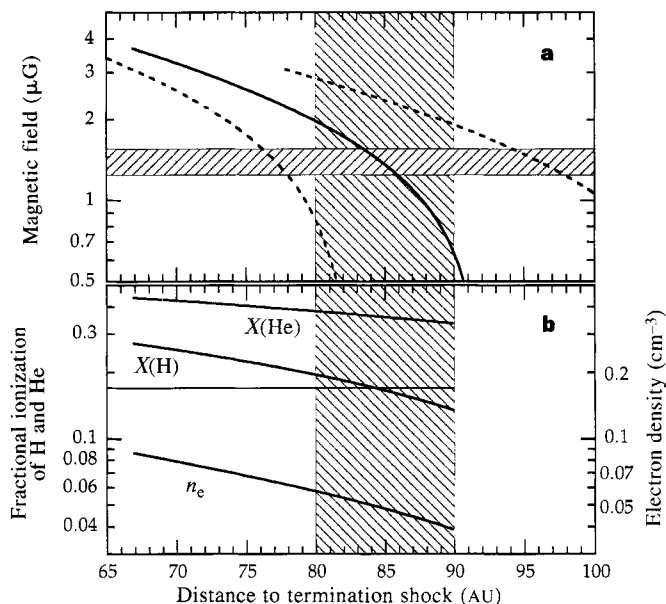


Figure 3 a. Variation of the calculated magnetic field strength, B , just beyond the bow shock with distance to the termination shock, R_{TS} . Mean (solid curve) and 1σ limits (dotted curves) are shown. $B^2/8\pi = \rho V_0^2 [(4/r) - 1 - 5r(2n_e kT)/(r+5)]$, where $V_0 = 26 \text{ km s}^{-1}$ is the relative speed of the Sun through the local cloud, r is the shock compression ratio, $\rho = m_p [n(\text{H}) + 4n(\text{He})]$, $n(\text{H}) = \phi(p)/r$ and $n(\text{He}) = [n(\text{H}) + n(\text{He})]/R2 - n(\text{He})$. The resulting stagnation pressure at the heliopause from the interstellar medium is then $P_{\text{LIC}} = [(7/8)(\rho V_0^2)] + [(3/8)(2n_e kT)] + (3B^2/8\pi) + P_H$, where $P_H = f n(\text{H}) [m_p V_0^2 + kT]$ is the pressure from that fraction $f = [1 - \exp(-\langle n(p) L \sigma \rangle)]$ of hydrogen atoms converted to 26 km s^{-1} protons in the filtration region by charge exchange. Magnetic field pressure and P_H contribute significantly to P_{LIC} . We neglect contributions from cosmic rays (as the observed gradients and thus the resulting force should be small) and from dust. The stagnation pressure at the heliopause from the solar wind, P_{SW} , is determined using the usual gas dynamic calculation, relating conditions just inside the termination shock, to those measured²³ at 49.1 AU by Voyager 2. We assume that in the distant heliosphere the solar wind expands uniformly but is decelerated slightly in the process of picking up interstellar atoms, neglect magnetic field pressure but include thermal pressure from pickup ions²³. $P_{\text{SW}} = (7/8)m_p [1.16 n_{\text{SW}}(p)] [49.1 U_{\text{SW}}/R_{\text{TS}}]^2 + (3/8)(m_p/7)(U_{\text{SW}}/R_{\text{TS}})^2 \beta_{\text{proc}} n(\text{H}_{\text{TS}})$, where $U_{\text{SW}} = V_{\text{SW}} [1 - (6/7)\sigma_p n(\text{H}_{\text{TS}} R_{\text{TS}}) / (1 - (6/7)\sigma_p n(\text{H}_{\text{TS}}) 49.1)]$ is the reduced solar-wind speed due to momentum loss to pickup hydrogen, $n_{\text{SW}}(p) = 0.0017 \text{ cm}^{-3}$, $V_{\text{SW}} = 534 \text{ km s}^{-1}$ is the 1996 averaged Voyager 2 solar-wind proton density and speed at 49.1 AU (ref. 23), the factor of 1.16 accounts for the solar-wind He/H of 4%, and $\beta_{\text{proc}} = 7 \times 10^{-7} \text{ s}^{-1}$. Assuming that there is no flow across or along the heliopause in the direction of motion of the Sun, the stagnation pressure on either side of the heliopause must be equal⁸. Parameter values from Table 1 are used to obtain solutions of the magnetic field as a function of R_{TS} . **b.** Average values of the electron density and ionization fractions of H and He calculated as a function of R_{TS} using values from Table 1 and computational steps given above. The shaded vertical region indicates the range of locations of the termination shock determined by Stone *et al.*¹⁹. The horizontal line is the calculated lower limit²² of X_{H} .

flow. The corresponding bow shock compression ratios range from 1.5 to 1.9.

The LIC is embedded in a region of tenuous, hot gas $\sim 80 \text{ pc}$ across^{1-3,17,20} with thermal pressure¹ between $\sim 1.2 \times 10^{-12}$ and $2.7 \times 10^{-12} \text{ dyn cm}^{-2}$. The LIC thermal pressure ($2n_e kT$) is at most $0.2 \times 10^{-12} \text{ dyn cm}^{-2}$, thus contradicting the basic assumption of thermal pressure balance in the interstellar medium^{1,2}. It has been suggested^{1,2} that pressure balance could be restored by adding sufficient LIC magnetic pressure. But this requires a minimum B of 5–7 μG , which is clearly inconsistent with our upper limit of 3 μG .

Could we substantially increase our upper limit on B of 3 μG by assuming subsonic conditions with no bow shock? For subsonic flow the pressure from the interstellar medium at the heliopause becomes⁸ $P_{\text{LIC}} = 0.5 \rho V_0^2 + 2n_e kT + 3B^2/8\pi + P_H$ (see Fig. 3a legend for definition of symbols). Using values in Table 1 with a compression ratio of 1 (no bow shock) we find that $B > 4.3 \mu\text{G}$ will push the average location of the termination shock below the minimum of 80 AU (ref. 19) to 70 AU. At that distance Voyager 1, now at 66 AU, would have occasionally intercepted the termination shock because of its $\pm(3-5) \text{ AU}$ excursions from its average position (in response to the $\pm(10-15)\%$ changes in the solar-wind ram pressure typically observed²³ over several months).

We now consider how to reconcile our extreme upper limit of $B = 4.3 \mu\text{G}$ with the minimum field of 5 μG required to achieve pressure balance with the hot cloud. One can abandon the idea that the magnetic field is uniform throughout the local cloud. Our results apply only to a very limited region of the LIC, within a few thousand AU of the Sun. The local cloud is likely to be complex and inhomogeneous³. Thus it is not unreasonable to imagine that the LIC field near the Sun is weak ($\sim 1-2 \mu\text{G}$) and becomes much stronger ($\sim 5-7 \mu\text{G}$) near its interface with the hot cloud. On the other hand, if it is assumed that the field is uniform but small ($\sim 1 \mu\text{G}$) it is still just possible to achieve pressure balance with the hot cloud and have the termination shock at 80 AU by postulating the existence of substantial ($\sim 10^{-12} \text{ dyn cm}^{-2}$) non-thermal pressure, perhaps residing in an undetected high-energy tail of the LIC plasma. The expected detection of the termination shock by Voyager 1 (within perhaps the next 5 years) should allow us to derive more accurate values for the magnetic field in the very local interstellar cloud. □

Received 9 September 1996; accepted 4 February 1997.

1. Bowyer, S., Lieu, R., Sidher, S. D., Lampton, M. & Kunde, J. *Nature* **375**, 212–214 (1995).
2. Cox, D. *Nature* **375**, 185–186 (1995).
3. Cox, D. P. & Reynolds, R. J. *Annu. Rev. Astron. Astrophys.* **25**, 303–344 (1987).
4. Möbius, E. *et al.* *Nature* **318**, 426–429 (1985).
5. Gloeckler, G. *et al.* *Science* **261**, 70–73 (1993).
6. Gloeckler, G. & Geiss, J. *Nature* **381**, 210–212 (1996).
7. Gloeckler, G. in *Proc. 1st ISSI Workshop* (eds von Steiger, R., Lallement, R. & Lee, M. A.) *Space Sci. Rev.* **78**, 335–346 (1996).
8. Holzer, T. E. *Annu. Rev. Astron. Astrophys.* **27**, 199–234 (1989).
9. Baranov, V. B. & Malama, Y. G. *J. Geophys. Res.* **100**, 755–761 (1995).
10. Williams, L. L., Hall, D. T., Pauls, H. L. & Zank, G. P. *Astrophys. J.* (submitted).
11. Steinolfson, R. S. & Gurnett, D. A. *Geophys. Res. Lett.* **22**, 651–654 (1985).
12. Gloeckler, G. *et al.* *Astron. Astrophys. Suppl. Ser.* **92**, 267–289 (1992).
13. Witte, M., Rosenbauer, H., Banaskiewicz, M. & Fahr, H. *J. Adv. Space Res.* **13**, 121–126 (1993).
14. Thomas, G. E. *Annu. Rev. Earth Planet. Sci.* **6**, 173–204 (1978).
15. Quémeneris, E., Bertaux, J.-L., Sandel, B. R. & Lallement, R. *Astron. Astrophys.* **290**, 941–954 (1996).
16. Valleria, J. in *Proc. 1st ISSI Workshop* (eds von Steiger, R., Lallement, R. & Lee, M. A.) *Space Sci. Rev.* **78**, 277–288 (1996).
17. Frisch, P. C. *Space Sci. Rev.* **72**, 499–592 (1995).
18. Siscoe, G. L. in *Solar System Magnetohydrodynamics: Concepts and Basic Equations in Solar Terrestrial Physics* (eds Carovillano, R. L. & Forbes, J. M.) 74–76 (Reidel, Dordrecht, 1983).
19. Stone, E. C., Cummings, A. C. & Webber, W. R. *J. Geophys. Res.* **101**, 11017–11025 (1996).
20. Frisch, P. C. in *Cosmic Winds and the Heliosphere* (eds Jokipii, J. R., Sonett, C. P. & Giampapa, M. S.) (Univ. Arizona Press, Tucson, in the press).
21. Dupuis, J., Vennes, S., Bowyer, S., Pradhan, A. K. & Thejll, P. *Astrophys. J.* **455**, 574–589 (1995).
22. Cheng, K.-P. & Bruhweiler, F. C. *Astrophys. J.* **364**, 573–581 (1990).
23. Burlaga, L. F., Ness, N. F., Belcher, J. W., Lazarus, A. J. & Richardson, J. D. in *Proc. 1st ISSI Workshop* (eds von Steiger, R., Lallement, R. & Lee, M. A.) *Space Sci. Rev.* **78**, 33–42 (1996).
24. Geiss, J. & Witte, M. in *Proc. 1st ISSI Workshop* (eds von Steiger, R., Lallement, R. & Lee, M. A.) *Space Sci. Rev.* **78**, 229–238 (1996).
25. Anders, E. & Grevesse, N. *Geochim. Cosmochim. Acta* **53**, 197–214 (1988).
26. Fite, W. L., Smith, A. C. H. & Stebbings, R. F. *Proc. R. Soc. Lond. A* **268**, 527–534 (1962).
27. Lallement, R., Bertaux, J.-L. & Clarke, J. T. *Science* **260**, 1095–1098 (1993).
28. Janev, R. K., Langer, W. D., Evans, K. Jr & Post, D. E. *Elementary Processes in Hydrogen-Helium Plasmas* (Cross Sections and Reaction Rate Coefficients) (Springer, Berlin, 1987).

29. Gloeckler, G., Schwadron, N. A., Fisk, L. A. & Geiss, J. *Geophys. Res. Lett.* **22**, 2665–2668 (1995).
 30. Fisk, L. A., Schwadron, N. A. & Gloeckler, G. *Geophys. Res. Lett.* **24**, 93–96 (1997).
 31. Rucinski, D. *et al.* *Proc. 1st ISSI Workshop* (eds von Steiger, R., Lallemand, R. & Lee, M. A.) *Space Sci. Rev.* **78**, 73–84 (1996).

Acknowledgements. The SWICS instrument was developed by a collaboration of the universities of Maryland, Bern and Braunschweig, and the Max-Planck-Institut für Aeronomie. We thank C. Gloeckler for her help with data reduction, and P. C. Frisch, E. Möbius and D. A. Gurnett for discussions. This work was supported by NASA/JPL and the Swiss National Science Foundation.

Correspondence should be addressed to G.G. (e-mail: gloeckler@umdsp.umd.edu).

Storage of hydrogen in single-walled carbon nanotubes

A. C. Dillon*, K. M. Jones*, T. A. Bekkedahl*, C. H. Kiang†, D. S. Bethune† & M. J. Heben*

* National Renewable Energy Laboratory, 1617 Cole Boulevard, Golden, Colorado 80401-3393, USA

† IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120-6099, USA

Pores of molecular dimensions can adsorb large quantities of gases owing to the enhanced density of the adsorbed material inside the pores¹, a consequence of the attractive potential of the pore walls. Pederson and Broughton have suggested² that carbon nanotubes, which have diameters of typically a few nanometres, should be able to draw up liquids by capillarity, and this effect has been seen for low-surface-tension liquids in large-diameter, multi-walled nanotubes³. Here we show that a gas can condense to high density inside narrow, single-walled nanotubes (SWNTs). Temperature-programmed desorption spectroscopy shows that hydrogen will condense inside SWNTs under conditions that do not induce adsorption within a standard mesoporous activated carbon. The very high hydrogen uptake in these materials suggests that they might be effective as a hydrogen-storage material for fuel-cell electric vehicles.

Soots containing SWNTs were synthesized by co-evaporation of cobalt and graphite in an electric arc⁴. Amorphous carbon-coated fibres several micrometres in length were observed in the soots by transmission electron microscopy (TEM). Fibres typically consisted of 7–14 bundled SWNTs which were individually ~12 Å in diameter⁵. Cobalt nanoparticles, 5–50 nm in diameter, were also present, embedded in amorphous carbon. Electron microprobe analysis measured cobalt contents of ~20 wt%. The remainder of each sample consisted of amorphous carbon and planar graphitic fractions. Activated carbon (AC) was prepared from pitch precursors, activated by KOH, and exhibited a narrow peak in the pore-volume distribution of 30 Å with a full-width at half-maximum of ~20 Å. (Material and pore analysis was supplied by Spectracorp Ltd.) Extensive TEM examinations showed the AC to be similar in structure and overall morphology to the arc-generated soots except for the absence of SWNTs and cobalt nanoparticles.

The adsorption of H₂ on SWNT soots and AC was probed with temperature programmed desorption (TPD) spectroscopy in an ultra-high-vacuum chamber equipped with a liquid-nitrogen-cooled cryostat and a mass spectrometer. Samples weighing ~1 mg were contained in a platinum foil packet with pinholes for gas diffusion. The packet temperature was measured with a thermocouple and controlled by resistive heating. TPD experiments were performed at 1 K s⁻¹, and standard H₂ exposures were carried out at 300 torr for 10 minutes at 273 K followed by 3 minutes at 133 K

except where noted. The samples were cooled to 90 K while the chamber was evacuated to $< 5 \times 10^{-8}$ torr before TPD.

H₂ desorbs from as-prepared SWNT soots (Fig. 1a) and AC (Fig. 1b) within the same temperature range, but with differing intensities. The signals are peaked just above 133 K because the H₂ pressure was reduced as the sample was cooled from 133 K to 90 K. Planar graphite cannot stabilize H₂ under these exposure conditions⁶, so desorption must be from nanoporous environments⁶ associated with the high-surface-area amorphous carbon fractions present in both materials. The signal from the SWNT samples (Fig. 1a) is ~10 times as great as the signal from AC (Fig. 1b), consistent with the higher amorphous carbon content (as

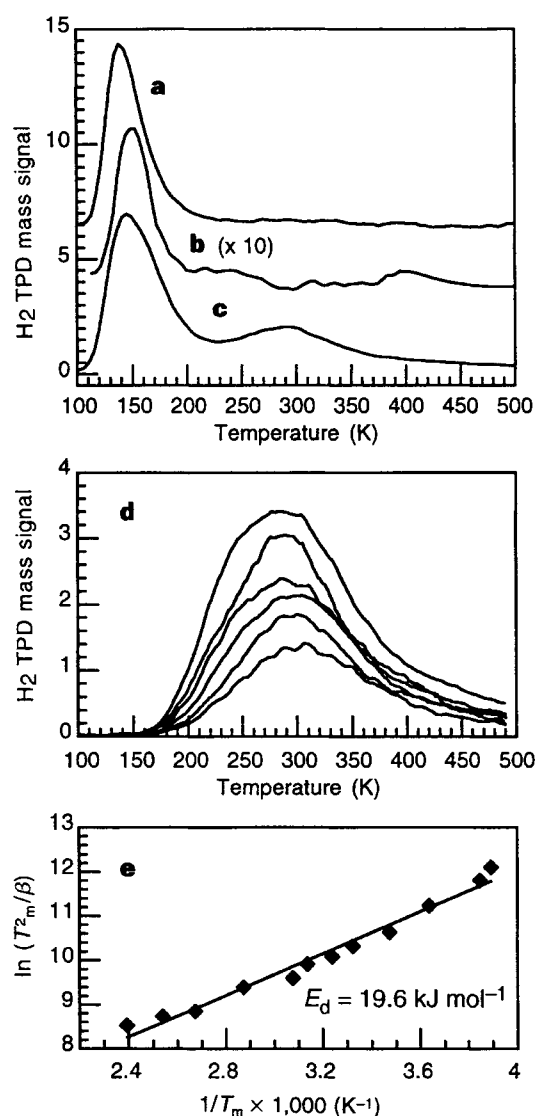


Figure 1 Temperature programmed desorption data. **a**, TPD spectrum from as-produced SWNT sample after standard hydrogen exposure. **b**, TPD spectrum from activated carbon sample, magnified 10 times, after standard hydrogen exposure. **c**, TPD spectrum from SWNT sample after heating in vacuum to 970 K and standard hydrogen exposure. **d**, Hydrogen desorption signals after exposures that populated only the high-temperature sites. Coverages range from 0.3 to saturation. **e**, Plot of $\ln(T_m^2/\beta)$ against $1/T_m$ with β varying from 0.3 to 30 K s⁻¹. The activation energy for desorption, E_d , is the slope of the line (19.6 kJ mol⁻¹).

observed by TEM) in SWNT samples. Storage of H_2 on AC materials at temperatures below 150 K has been investigated before⁷ but is unattractive because cooled compressed gas containers perform equally well with or without the adsorbent^{7,8}. We note that all TPD peaks presented here vary by $\pm 20\%$ in height and $\pm 5^\circ$ in position from sample to sample, but are highly reproducible from run to run on a given sample.

A high-temperature TPD peak representing a population of structurally unique sites is found at 288 K after SWNT soots are heated in vacuum to 970 K at 1 K s^{-1} (Fig. 1c). The adsorbed H_2 is stable at 223 K while in contact with vacuum and does not redistribute to other sites during evacuation or cooling. Figure 1d shows the TPD spectra when the high-temperature sites were selectively populated by evacuating the chamber with a sample temperature of 223 K after a 10-minute H_2 exposure at 273 K. Variation of the hydrogen pressure between 25 and 300 torr produced coverages between ~ 0.3 and saturation. The shift in peak temperature from 299 to 285 K with decreasing coverage is explained by first-order desorption of physisorbed H_2 with either coverage-dependent desorption energies⁹ or equilibrium readsorption¹⁰. Shifts of 100 K or more are expected for the second-order desorption kinetics which would characterize the desorption of chemisorbed H_2 (ref. 9). First-order desorption is described by $\ln(T_m^2/\beta) = E_d/RT_m$ where T_m is the temperature at the peak maximum, E_d is the desorption activation energy, β is the heating rate and R is the universal gas constant⁹. T_m was measured for β between 0.3 and 30 K s^{-1} after only the high-temperature sites were populated at 300 torr. A plot of $\ln(T_m^2/\beta)$ against $1/T_m$ yields a straight line (Fig. 1e), confirming that H_2 is physisorbed in the high-temperature sites. The desorption

activation energy, which is equivalent to the heat of adsorption in the case of physisorption⁹, is 19.6 kJ mol^{-1} .

The strongly bound physisorbed H_2 that we observe is consistent with adsorption within the cavities of SWNTs^{1,2,11}, indicating that the tube interiors become accessible after heating in vacuum. Multi-wall carbon nanotubes are opened by reaction with either O_2 (ref. 12) or CO_2 (ref. 13) because of the local strain and incorporated pentagons in the tube caps. The caps of small-diameter SWNTs should be susceptible to oxidation for the same reasons. Infrared absorption and mass spectroscopies show H_2O and CO_2 are generated by the decomposition of surface groups upon heating in vacuum. The production of H_2O and CO_2 decreases at temperatures approaching 700 K while CO generation increases. Although it is difficult to observe by TEM, we believe that surface carbon and SWNT caps are removed during this process. Arguably, heating in vacuum may simply produce a clean substrate for subsequent H_2 adsorption. However, H_2 uptake can be increased by a factor of ~ 3 after gentle oxidation with 1 torr H_2O . Reactive carbonaceous fractions are converted primarily to CO during the oxidation process, and additional SWNTs are evidently opened while the sample weight is decreased to $\sim 1/3$ of its initial value. Unfortunately, the degree to which SWNTs can be purified with this approach by removing amorphous carbon is limited, as SWNTs are also consumed with further oxidation. It is important to note that high-temperature desorption of H_2 is never observed from AC samples, or from arc-generated soots produced without the cobalt catalyst (which therefore contain no SWNTs).

The cobalt nanoparticles must also be considered as a possible source for the high-temperature H_2 desorption. However, H_2 TPD studies on pure cobalt single crystals¹⁴ and polycrystalline samples¹⁵ find TPD peaks at different temperatures, and observe second-order desorption kinetics characterizing recombinative desorption of chemisorbed H_2 (refs 14, 15). Furthermore, the amount of H_2 observed during TPD is inconsistent with either a cobalt surface or bulk hydride. Electron diffraction of cobalt nanoparticles in SWNT soots show face-centred-cubic patterns⁴ initially, and a blurred superposition of hexagonal Co_3O_4 and cobalt patterns after heating in vacuum. The composite pattern is expected for cobalt metal particles covered with a thin oxide, and this coating may explain the lack of H_2 uptake by the cobalt nanoparticles.

SWNTs are the only other component unique to the soots, and are apparently responsible for stabilizing H_2 by physisorption at unusually high temperatures. Planar graphite adsorbs a complete monolayer of H_2 with a constant isosteric heat of adsorption of $\sim 4\text{ kJ mol}^{-1}$ (ref. 5). Heats as high as 12 kJ mol^{-1} have been observed with certain ACs at very low coverages, but decrease to $\sim 4\text{ kJ mol}^{-1}$ at higher coverages⁷. Nanometre-sized pores with high binding energies for H_2 are evidently present in AC, but constitute only a small fraction of the total porous volume. The energy of 19.6 kJ mol^{-1} measured with SWNT materials is evidence for H_2 localized in the small diameter tubes^{1,2,11}. The high-temperature sites cannot be occupied in 10 minutes at 133 K as expected for slow diffusion along the internal SWNT surfaces.

The amount of H_2 adsorbed in the high-temperature peak for a sample that had been progressively oxidized in H_2O was $\sim 0.01\text{ wt\%}$. Graphical integration methods applied to TEM micrographs of this sample indicated a SWNT density ranging from 0.1 to 0.2 wt% depending on the average number of tubes estimated per bundle (7 or 14, respectively). Thus, the gravimetric storage density per SWNT ranges from ~ 5 to 10 wt%. These values are 2.5 to 5 times greater than expected for filling tubes with close-packed H_2 at a nearest-neighbour distance of $3.51\text{ \AA}$ (ref. 16), and a closest approach to the tube wall of $2.95\text{ \AA}$ (ref. 5). The difference may be rationalized if SWNTs were depleted on the TEM sample grids relative to their density in the soots. Alternatively, H_2 could be packed within the tubes at higher density than assumed if H_2 - H_2 repulsive interactions were screened^{2,16}, or additional H_2 might be

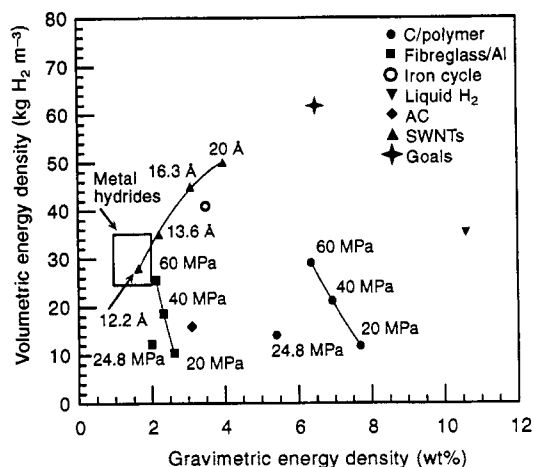


Figure 2 Installed energy densities for several vehicular hydrogen storage technologies. Lines on the plot are a guide to the eye. Calculations²¹ for carbon/polymer and fibreglass/aluminium cylinders at 20, 40 and 60 MPa used performance factors of 3.5×10^5 and $1.1 \times 10^5\text{ m}^2\text{ s}^{-2}$, a ratio of outer/inner volume of 1.2 and 1.4, respectively, and a burst pressure three times the operating pressure. Data at 24.8 MPa are for currently manufactured containers (Structural Composites Industries and EDO Corporation). Required brackets and pressure regulators are not included. The area for metal hydrides encloses representative data^{18,19} and considers the system weight and volume for both low- and high-dissociation-temperature materials. Data for liquid H_2 are from ref. 17, and those for adsorption on activated carbon at 155 K and 6.9 MPa are from ref. 7. The iron cycle uses on-board H_2O to generate H_2 by reaction with Fe, and requires Fe_3O_4 to be reduced to Fe off the vehicle¹⁹. The SWNT data are predicted for individual tubes assembled into hexagonally packed bundles assuming H_2 packing as described in the text. The points are labelled with diameters which, from smallest to largest, correspond to (9,9), (10,10), (12,12), and (15,15) 'armchair' tubes²². A low-pressure container with a weight equal to the SWNT weight divided by 3.5, and an outer volume 1.075 times the inner volume, is included.

adsorbed by the exterior surfaces of SWNTs and/or the interstitial spaces between bundled tubes.

A vehicle powered by a fuel cell would require ~ 3.1 kg of H_2 for a 500 km range¹⁷. This amount of H_2 stored in the weight and volume of a typical petrol tank requires system densities approaching 6.5 wt% and 62 kg H_2 m⁻³ (ref. 17). Figure 2 shows that no storage technology is currently capable of meeting these goals. SWNTs with diameters of 16.3 Å and 20 Å would come close to the target densities and operate near room temperatures if modest H_2 overpressures compensated for the lower heats of adsorption expected in the larger cavities. These materials would have high energy storage efficiencies as they would operate at or near ambient temperatures and pressures. In contrast, 25 to 45% of the energy content in liquefied H_2 is required for liquefaction¹⁸, and $\sim 9\%$ of the stored energy is needed for compression of H_2 to 20 MPa (ref. 17). For catalytic generation of H_2 by the reaction of H_2O with iron, temperatures in excess of 250 °C are required^{17,19}.

The effect of hydrogen over-pressure on the stability of adsorbed H_2 has not yet been investigated and remains an important question. The high-purity 13.8-Å diameter SWNT samples that have recently been produced by laser vaporization²⁰ should be interesting candidates for evaluation. The temperature and pressure requirements for H_2 adsorption and desorption, and the kinetics for charging and discharging, are expected to be a function of nanotube diameter and aspect ratio. Control of these parameters coupled with improvements in production, purification and alignment of SWNTs may lead to a new H_2 storage technology for hydrogen-fuelled vehicles with superior performance to currently available options. □

Received 4 November 1996; accepted 10 February 1997.

- Gregg, S. J. & Sing, K. S. W. *Adsorption, Surface Area and Porosity* (Academic, London, 1982).
- Pederson, M. R. & Broughton, J. Q. Nanocapillarity in fullerene tubules. *Phys. Rev. Lett.* **69**, 2689–2692 (1992).
- Dujardin, E., Ebbesen, T. W., Hiura, H. & Tanigaki, K. Capillarity and wetting of carbon nanotubes. *Science* **265**, 1850–1852 (1994).
- Bethune, D. S. *et al.* Cobalt-catalysed growth of carbon nanotubes with single-atomic-layer walls. *Nature* **363**, 605–607 (1993).
- Pace, E. L. & Siebert, A. R. Heat of adsorption of parahydrogen and orthodeuterium on graphon. *J. Phys. Chem.* **63**, 1398–1400 (1959).
- Bandos, T. J., Jagiello, J., Amankwah, K. A. G. & Schwarz, J. A. Chemical and structural properties of clay minerals modified by inorganic and organic material. *Clay Miner.* **27**, 435–444 (1992).
- Schwarz, J. A. *Final Report for the Tasks XC-1-1108-1 and XAE-3-13346-01* (National Renewable Energy Laboratory, Golden, Colorado, 1994).
- Dillon, A. C., Bekkedahl, T. A., Cahill, A. F., Jones, K. M. & Heben, M. J. Carbon nanotube materials for hydrogen storage. *Proc. 1995 U.S. DOE Hydrogen Program Review* 521–541 (National Renewable Energy Laboratory, Golden, Colorado, 1995).
- Madix, R. J. The application of flash desorption spectroscopy to chemical reactions on surfaces: Temperature programmed reaction spectroscopy. *Chemistry and Physics of Solid Surfaces* (ed. Vanselow, R.) 63–72 (CRC, Boca Raton, 1979).
- Ibok, E. E. & Ollis, D. F. Temperature programmed desorption from porous catalysts: Shape index analysis. *J. Catal.* **66**, 391–400 (1980).
- Peterson, B. K. & Gubbins, K. E. Phase transitions in a cylindrical pore: Grand canonical Monte Carlo, mean field theory, and the Kelvin equation. *Molec. Phys.* **62**, 215–226 (1987).
- Ajayan, P. M. *et al.* Opening carbon nanotubes with oxygen and implications for filling. *Nature* **362**, 522–525 (1993).
- Tsang, S. C., Harris, P. J. F. & Green, M. L. H. Thinning and opening of carbon nanotubes by oxidation using carbon dioxide. *Nature* **362**, 520–522 (1993).
- Ernst, K. H., Schwarz, E. & Christmann, K. The interaction of hydrogen with a cobalt (1010) surface. *J. Chem. Phys.* **101**, 5388–5401 (1994).
- Lisowski, W. The kinetics of the low-temperature hydrogen interaction with polycrystalline cobalt films. *Appl. Surf. Sci.* **37**, 272–282 (1989).
- Nielsen, M., McTague, J. P. & Ellenson, W. Adsorbed layers of D_2 , H_2 , O_2 , and 4He on graphite studied by neutron scattering. *J. Phys. Chem.* **81**, 4110–4118 (1977).
- DeLuchi, M. *Hydrogen Fuel-Cell Vehicles* (Institute of Transportation Studies, Univ. California, Davis, 1992).
- DeLuchi, M. A. Hydrogen vehicles: An evaluation of fuel storage, performance, safety, environmental impacts, and cost. *Int. J. Hydrogen Energy* **14**, 81–130 (1989).
- T-Raissi, A. & Sadhu, A. Systems study of metal hydride storage requirements. *Proc. 1994 DOE/NREL Hydrogen Program Review* 85–106 (National Renewable Energy Laboratory, Golden, Colorado, 1994).
- Thess, A. *et al.* Crystalline ropes of metallic carbon nanotubes. *Science* **273**, 483–487 (1996).
- Gordon, R. Composite pressure vessels for gaseous hydrogen-powered vehicles. *Hydrogen Energy Progress V* (eds Vezirgil, T. N., Taylor, J. B.) 1225–1236 (Pergamon, New York, 1984).
- Dresselhaus, M. S., Dresselhaus, G. & Saito, R. C_{60} related tubules. *Solid State Commun.* **84**, 201–205 (1992).

Acknowledgements. We thank Jeremy Broughton, J. Karl Johnson and Al Czanderna for technical discussions. We also thank Spectracorp Limited for activated carbon samples, and A. Mason for microprobe measurements. This work was funded by the US Department of Energy Hydrogen Program.

Correspondence and requests for materials should be addressed to M.J.H. (e-mail: mikhe@nrel.gov).

Spontaneous stratification in granular mixtures

Hernán A. Makse*, Shlomo Havlin†‡, Peter R. King‡ & H. Eugene Stanley*

* Center for Polymer Studies and Physics Department, Boston University, Boston, Massachusetts 02215, USA

† Minerva Center and Department of Physics, Bar-Ilan University, Ramat Gan, Israel

‡ BP Exploration Operating Company Ltd, Sunbury-on-Thames, Middlesex, TW16 7LN, UK

Granular materials^{1–5} segregate according to grain size when exposed to periodic perturbations such as vibrations^{6–12}. Moreover, mixtures of grains of different sizes can also spontaneously segregate in the absence of external perturbations: when such a mixture is simply poured onto a pile, the large grains are more likely to be found near the base, while the small grains are more likely to be near the top^{13–20}. Here we report another size-separation effect, which arises when we pour a granular mixture between two vertical plates: the mixture spontaneously stratifies into alternating layers of small and large grains whenever the large grains have larger angle of repose than the small grains. We find only spontaneous segregation, without stratification, when the large grains have smaller angle of repose than the small grains. The stratification is related to the occurrence of avalanches: during each avalanche, the grains separate into a pair of static layers, with the small grains forming a sublayer underneath the layer of large grains.

Our experimental system consists of a vertical ‘quasi-two-dimensional’ cell with a gap of 5 mm separating two transparent plates (made of Plexiglass or glass) measuring 300 mm × 200 mm (Fig. 1a). To avoid the effects of electrostatic interactions between the grains and the wall, the wall is cleaned with an antistatic cleaner.

In a first series of experiments, we closed the left edge of the cell leaving the right edge free, and we poured, near the left edge, an equal-volume mixture of white glass beads (mean size 0.27 mm, spherical shape, repose angle 26°), and red sugar crystals (typical size 0.8 mm, cubic shape, repose angle 39°). Figure 1a shows the result of the first series of experiments. We note two features: (1) Spontaneous stratification. We see the formation of alternating layers consisting of small and large grains—with a ‘wavelength’ of ~ 1.2 cm. (2) Spontaneous segregation. We find that the smaller grains segregate near the left edge and the larger grains segregate furthest from it and near the base^{13–20}.

In a second series of experiments, we confirmed the results of these initial experiments by testing for stratification and segregation using a mixture of grains of same density, consisting of fine sand (typical size 0.4 mm) and coarse sand (typical size 1 mm), suggesting that the density of the grains may not play an important role in stratification.

In all the above experiments we used mixtures composed of two types of grain with different shape, and therefore with different angles of repose. In particular we obtained stratification (plus segregation) when we used larger cubic grains and smaller spherical grains: the angle of repose of the large species was then larger than the angle of repose of the small species. Otherwise we obtained only segregation and not stratification when the large grains were less faceted than the small grains, that is, the large grains had smaller angle of repose than the small grains.

To confirm this, we performed a series of experiments using mixtures of irregular shaped sand grains (repose angle 35°, mean size 0.3 mm), and spherical glass beads (repose angle 26° smaller than the repose angle of the sand grains). We found that stratification

(plus segregation) occurred for two different experiments using spherical beads of size 0.07 mm and 0.11 mm (so that the larger grains had larger repose angle). In contrast, we obtained only segregation but not stratification for two experiments using spherical beads of size 0.55 mm and 0.77 mm (so that the larger grains had smaller repose angle). In all cases the segregation of grains occurred with the smaller grains being found near the left edge of the cell and the larger grains near the base of the cell. These results suggest that the phenomenon of segregation is always expected when pouring a granular mixture of grains of different sizes, no matter what are the values of the angles of repose of the species. However, the phenomenon of stratification is only expected when the large species have larger angle of repose than the small species.

Additionally, we performed a series of experiments (not shown) in which we found similar stratification by using different mixtures of differing size ratio between large and small grains (1.66, 2.1, 2.25, 3.25 and 6.66), suggesting that the phenomenon occurs for a broad regime of grain size ratios. We found a similar stratification when we doubled the gap between the vertical plates of the cell and simultaneously doubled the flow rate of grains.

We propose a physical mechanism responsible for the observed stratification that is related to the fact that not one but rather a pair of layers is formed in the course of each avalanche. When the flow of grains reaches the base of the pile, we find that the grains develop a profile characterized by a well-defined 'kink', at which the grains are stopped (Fig. 1b); but we find that the small grains stop first, so a pair of layers forms with the small grains underneath the large grains. As more grains are added, the kink appears to move upwards in the direction opposite to the flow of grains. Once the kink reaches the top, the pair of layers is complete and the cycle is then repeated: a new avalanche occurs, the kink develops, and a new pair of layers forms.

The 'wavelength' of a pair of layers λ can be determined by the mean value of the downward velocity v of the rolling grains during an avalanche, the upward velocity v' of the kink, and the thickness of the layer of rolling grains R_0 during the avalanche. If the volume of grains in an avalanche scales approximately as the volume of grains in a well formed kink, we predict $\lambda \approx R_0(v + v')/v'$, and we confirm this relation experimentally.

To test this physical mechanism by computer simulation we

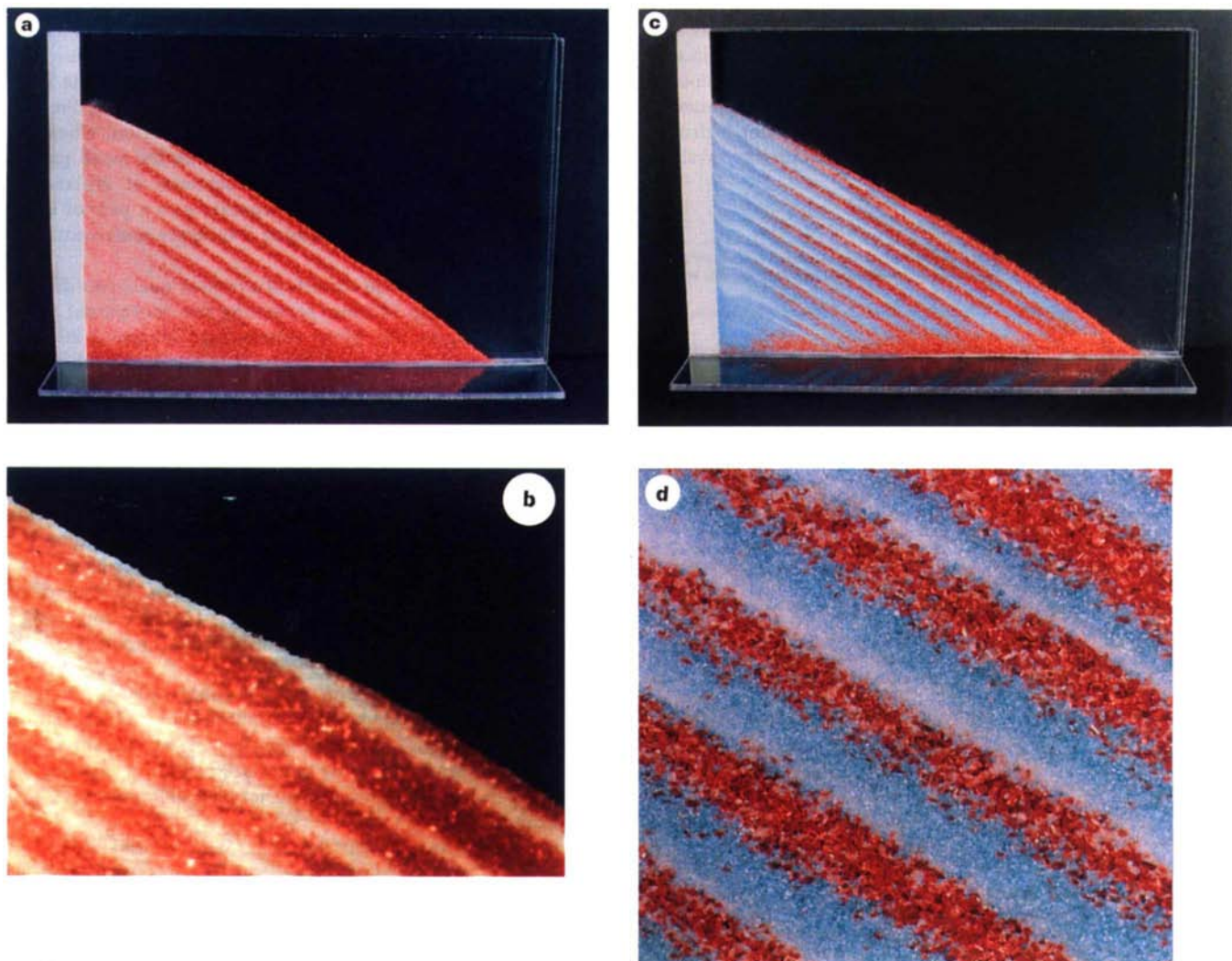


Figure 1 Experimental results. **a**, Typical result of the first series of experiments, showing the formation of successive layers of fine and coarse grains (here the white grains are glass beads of average diameter 0.27 mm, while the larger grains are sugar crystals of typical size 0.8 mm). We poured the equal-volume mixture near the left edge between two transparent vertical plates separated by a gap of 5 mm. We obtained stratification with a wavelength $\lambda \approx 1.2$ cm. **b**, Close-up photograph of the kink where the grains stop during an avalanche. The small

white grains stop first, and then the large red grains; hence the small grains form a sublayer underneath the large grains. **c**, Stratification obtained using a mixture of three different types of grains: nearly spherical glass beads (0.15 mm, angle of repose 26°), blue sand (0.4 mm, angle of repose 35°) and red sugar crystals (0.8 mm, angle of repose 39°). We notice the grading (from bottom to top) in a triplet of layers: small (white), medium (blue) and large (red) grains. **d**, Close-up photograph of the stratification pattern of **c**, field of view 40 mm $\times$ 40 mm.

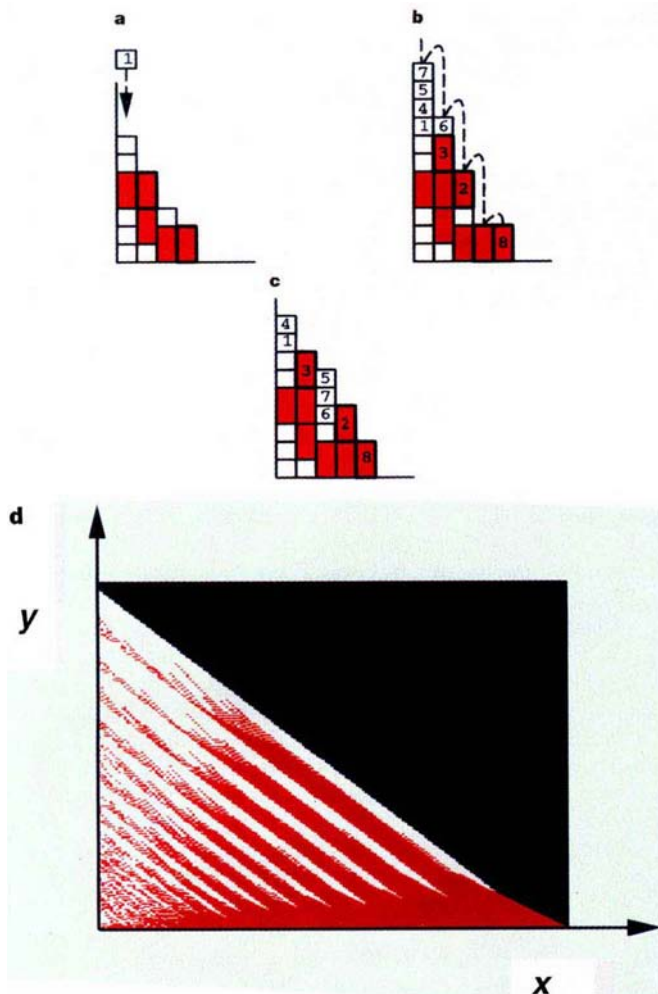


Figure 2 Results of modelling. **a**, The dynamics of the simplest 'mean field' model are illustrated by this example with two sizes $H_1 = 1$ (white) and $H_2 = 2$ (red), and threshold slopes $s_r = \tan\theta_r = 2$ and $s_m = \tan\theta_m = 3$. Suppose that, at a given instant, the sand pile is at the critical slope for repose s_r . To define the dynamical rules for the arriving grains, we consider the slopes $s_i = h_i - h_{i-1}$, where h_i denotes the height of the sand pile at coordinate i . We deposit a grain near the first column at the left edge of the lattice, where the actual column position is chosen from a narrow gaussian probability distribution centred at the wall edge. The non-zero width of this gaussian mirrors the fact that grains often bounce after reaching the pile. Newly arriving grains accumulate on the sand pile profile, following dynamics governed by the critical slope s_m ; thus a grain moves from the initial landing point at column i to column $i + 1$ if the slope $h_i - h_{i-1}$ is larger than s_m , then moves from column $i + 1$ to column $i + 2$ if $h_{i+1} - h_i > s_m$, and so forth. The grain stops at the first column k with $h_k - h_{k-1} \leq s_m$. Another grain is now added, and the same rules are followed. **b**, This entire process continues until a grain reaches the substrate at the furthest right column of the pile for first time (grain 8 in this figure). Now, as $s_i > s_m$ for all columns i , the sand pile has become 'unstable'. We note that s_i is calculated considering also the size of the rolling grain; as a result, the large grains more readily reach a slope that exceeds the two critical slopes s_r and s_m . **c**, We allow the sand pile to relax towards the repose slope s_r by moving each of the grains with slope larger than s_r to the nearest column satisfying $h_i - h_{i-1} \leq s_r$. Now the deposition starts again, and we iterate the algorithm until a large sand pile (of typically 10^5 grains) is formed. We can obtain stratification with constant 'wavelength' by stopping the accumulation process when a grain reaches for first time the column $i^{1/2}$, where i denotes the furthest-right column of the pile. **d**, Image obtained with the simplest 'mean field' model (for parameters $H_1 = 1$, $H_2 = 2$, $s_r = 4$ and $s_m = 5$); the smaller grains are white and the larger grains red. We find stratification and also reproduce the 'kink' mechanism explained in the text when we improve upon the simplest 'mean field' model by including four different angles of repose to take into account the fact that the angle of repose depends on the concentration of grains at the surface of the pile²³.

considered a mixture comprising small grains of width one pixel and of height H_1 , and large grains, also of width one pixel but of height $H_2 > H_1$. To generate an equal-volume mixture, we randomly 'dropped' a small grain with probability $p \equiv H_2/(H_1 + H_2)$, and 'dropped' a large grain with probability $1 - p$ (Fig. 2).

In critical phenomena, it is often useful to first develop a 'mean field' type model (in which, for example, spin orientation is determined by a macroscopic variable, the net magnetization), before devising models in which spin orientation is determined by the microscopic quantities such as the orientations of the other spins comprising the system. In this spirit, we focus first not on the 'microscopic' grain motions, but rather on the 'macroscopic' angle of the sand pile, whose value alternates in time between the maximum angle of stability θ_m which defines the onset of an avalanche, and the angle of repose θ_r which defines the end of the avalanche^{21,22}. Using this model (described in Fig. 2 legend), we found a morphology that displays both segregation and stratification.

In addition to the simplest 'mean field' approach, we developed a model²³ in which we treat the individual grain motion in accordance with microscopic rules that depend not on the macroscopic angle of the sand pile, but rather on the local angles formed between each grain and its neighbours. Specifically, the dynamics of the small and large rolling grains are governed by the critical angles of repose corresponding to the interactions between the rolling grain and the static grains of the sand-pile surface. This model incorporates the experimental fact that grains segregate because large grains roll down more easily on top of small grains than small grains on top of large grains (for rolling large grains on top of a surface of small grains, the surface appears smoother than for rolling small grains on top of a surface of large grains). Thus the model correctly predicts that the small grains form a sublayer beneath the large grains. We find stratification, as in the simplest 'mean field' model, and also find that the profile of the sandpile displays a kink at which rolling grains are stopped—just as in the experiment.

Next we tested the above principles by generalizing from two grain sizes to three. The experiment resulted in stratification with three layers, with the finest grains on the bottommost of each triplet of layers and the coarsest grains on the topmost layer (see Fig. 1c where the same experimental setup of Fig. 1a is used to obtain an alternation of three layers of grains of three different sizes: 0.15 mm, 0.4 mm and 0.8 mm; a close-up view of the stratification pattern of Fig. 1c is shown in Fig. 1d). We are now performing experiments using a continuum size distribution, as geological rock formations (which also display stratification) generally occur in the presence of a continuum distribution of grain size.

To provide another test of the proposed physical mechanism, we note that if both grains are spherical, there should be no stratification because the angles of repose of the large and small grains are then the same. We confirmed this expectation experimentally. The case of spherical grains was also studied by Williams^{24–26}, who seems to have seen a hint of the stratification effect. However, its origin in that case appears to have been an artefact, presumably because his grains were in fact of slightly different shapes.

We note that Boutreux and de Gennes have recently made considerable progress²⁷ in developing a general theoretical framework^{28,29} to treat the case of granular flows of two different grains. Their conclusions^{27–29} are consistent with the experiments presented here. □

Received 11 October 1995; accepted 10 February 1997.

1. Bagnold, R. A. *The Physics of Blown Sand and Desert Dunes* (Chapman & Hall, London, 1941).
2. Jaeger, H. M. & Nagel, S. R. Physics of the granular state. *Science* **255**, 1523–1531 (1992).
3. Herrmann, H. J. in *Disorder and Granular Media* (eds Bideaux, D. & Hansen, A.) 305–320 (North-Holland, Amsterdam, 1993).
4. Edwards, S. F. in *Granular Matter: An Interdisciplinary Approach* (ed. Mehta, A.) 121–140 (Springer, New York, 1994).
5. Wolf, D. E. in *Computational Physics: Selected Methods, Simple Exercises, Serious Applications* (eds Hoffman, K. H. & Schreiber, M.) 64–95 (Springer, Berlin, 1996).

6. Williams, J. C. The segregation of particulates materials. A review. *Powder Technol.* **15**, 245–251 (1976).
7. Rosato, A., Strandburg, K. J., Prinz, F. & Swendsen, R. H. Why the Brazil nuts are on top: size segregation of particulate matter by shaking. *Phys. Rev. Lett.* **58**, 1038–1040 (1987).
8. Gallas, J. A. C., Herrmann, H. J. & Sokolowski, S. Convection cells in vibrating granular media. *Phys. Rev. Lett.* **69**, 1371–1374 (1992).
9. Knight, J. B., Jaeger, H. M. & Nagel, S. R. Vibration-induced size separation in granular media: the convection connection. *Phys. Rev. Lett.* **70**, 3728–3731 (1993).
10. Zik, O., Levine, D., Lipson, S. G., Shtrikman, S. & Stavans, J. Rotationally induced segregation of granular materials. *Phys. Rev. Lett.* **73**, 644–647 (1994).
11. Clément, E., Rajchenbach, J. & Duran, J. Mixing of a granular material in a bi-dimensional rotating drum. *Europhys. Lett.* **30**, 7–12 (1995).
12. Cooke, W., Warr, S., Huntley, J. M. & Ball, R. C. Particle size segregation in a two-dimensional bed undergoing vertical vibration. *Phys. Rev. E* **53**, 2812–2822 (1996).
13. Brown, R. L. The fundamental principles of segregation. *J. Inst. Fuel* **13**, 15–19 (1939).
14. Bagnold, R. A. Experiments on a gravity-free dispersion of large solid spheres in a Newtonian fluid under shear. *Proc. R. Soc. Lond. A* **225**, 49–63 (1954).
15. Drahn, J. A. & Bridgwater, J. The mechanisms of free surface segregation. *Powder Technol.* **36**, 39–53 (1983).
16. Fayed, M. E. & Otten, L. (eds) *Handbook of Powder Science and Technology* 428–433 (Van Nostrand Reinhold, New York, 1984).
17. Savage, S. B. in *Developments in Engineering Mechanics* (ed. Selvadurai, A. P. S.) 347–363 (Elsevier, Amsterdam, 1987).
18. Savage, S. B. & Lun, C. K. Particle size segregation in inclined chute flow of dry cohesionless granular solids. *J. Fluid Mech.* **189**, 311–335 (1988).
19. Savage, S. B. in *Theoretical and Applied Mechanics* (eds Germain, P., Piau, M. & Caillerie, D.) 241–266 (Elsevier, Amsterdam, 1989).
20. Meakin, P. A simple two-dimensional model for particle segregation. *Physica A* **163**, 733–746 (1990).
21. Bagnold, R. A. The shearing and dilation of dry sand and the 'singing' mechanism. *Proc. R. Soc. Lond. A* **295**, 219–232 (1966).
22. Jaeger, H. M., Liu, C.-H. & Nagel, S. R. Relaxation at the angle of repose. *Phys. Rev. Lett.* **62**, 40–43 (1989).
23. Makse, H. A., Cizeau, P. & Stanley, H. E. Possible stratification mechanism in granular mixtures. *Phys. Rev. Lett.* (submitted).
24. Williams, J. C. The segregation of powders and granular materials. *Univ. Sheffield Fuel Soc. J.* **14**, 29–34 (1963).
25. Williams, J. C. The mixing of dry powders. *Powder Technol.* **2**, 13–20 (1968).
26. Allen, J. R. L. *Sedimentary Structures: their Character and Physical Basis* (Elsevier, Amsterdam, 1982).
27. Bouteux, T. & de Gennes, P.-G. Surface flows of granular mixtures: I. General principles and minimal model. *J. Phys. I France* **6**, 1295–1304 (1996).
28. Bouchaud, J.-P., Cates, M. E., Prakash, J. R. & Edwards, S. F. Hysteresis and metastability in a continuum sandpile model. *Phys. Rev. Lett.* **74**, 1982–1985 (1995).
29. de Gennes, P.-G. Dynamique superficielle d'un matériau granulaire. *C. R. Acad. Sci.* **321** [IIB], 501–506 (1995).

Acknowledgements. We thank T. Bouteux, G. Davies, P.-G. de Gennes, H. J. Herrmann and S. Tomassone for valuable discussions. P. Cizeau for collaboration in stages of this work, K. Papworth for technical assistance with the photographs, and BP for financial support.

Correspondence should be addressed to H.A.M. (hmakse@miranda.bu.edu).

Closure of the Central American Isthmus and its effect on deep-water formation in the North Atlantic

Kevin W. Burton[†], Hong-Fei Ling[‡] & R. Keith O'Nions^{*}

^{*} Department of Earth Sciences, University of Oxford, Parks Road, Oxford OX1 3PR, UK

[‡] Department of Earth Sciences, University of Nanjing, Nanjing 210008, China

Modern ocean thermohaline-driven circulation influences global climate by transporting heat to high latitudes^{1,2} and by affecting the exchange of CO₂ between ocean and atmosphere³. North Atlantic Deep Water (NADW) plays a key role in this circulation, and Quaternary climate cycles have been linked to changes in NADW flow⁴. General circulation model simulations indicate that before closure, some 3–4 million years ago, of the Central American Isthmus—the narrow strip of land linking North and South America—the direct flow of low-salinity water from the Pacific to the Atlantic Ocean would have led to a smaller NADW flow^{5,6}. Sedimentation patterns⁷ and nutrient proxies^{8–11} support these model results by indicating an increase in NADW flow around the

time of isthmus closure, but these records do not allow changes in different NADW sources to be distinguished, and the overall effect of closure on global ocean circulation is poorly known. Here we present Nd, Pb and Sr isotope records preserved by a hydrogenous ferromanganese crust from the NADW flow-path in the western North Atlantic Ocean. These records indicate that the isotopic signal associated with NADW strengthened around 3–4 million years ago showing that deep water that formed in the Labrador Sea made a gradually increasing contribution to NADW flow. These data, taken together with those from the central Pacific Ocean¹², indicate an increased NADW flow since isthmus closure, and suggest that the closure established today's general pattern of ocean circulation.

North Atlantic Deep Water possesses a distinct physical, chemical and isotopic signature (see, for example, refs 4, 13–15). In particular, the sources of NADW impart an extremely unradiogenic Nd isotope composition ($\epsilon_{\text{Nd}} = -13.5 \pm 0.5$ (ref. 15), where ϵ_{Nd} is the measured $^{143}\text{Nd}/^{144}\text{Nd}$ ratio relative to the chondrite uniform reservoir¹⁶; see Fig. 2a legend) which reflects local input through erosion of old continental material, consistent with the age of the surrounding land masses. By comparison, Antarctic Bottom Water (AABW), the only other major source of deep water in the present-day Atlantic, has a much more radiogenic Nd isotope composition ($\epsilon_{\text{Nd}} = -8.6 \pm 0.5$)¹⁷. These differences in isotopic composition ($\Delta\epsilon_{\text{Nd}} = 4.9 \pm 1.0$) are much larger than the present-day measurement precision for the $^{143}\text{Nd}/^{144}\text{Nd}$ ratio (± 20 p.p.m. (2σ) $\approx 0.20 \epsilon_{\text{Nd}}$). Therefore, in principle, if an appropriate record can be found, it should be possible to detect the variations in seawater chemistry that would occur as a consequence of changes in deep-water circulation.

Hydrogenous ferromanganese crusts grow by direct accumulation of metal oxides from sea water, and thus potentially preserve a record of the chemistry of the sea water from which they grow. Here we present Nd, Pb and Sr isotope data for an Fe–Mn crust (BM1969.05) from the San Pablo seamount in the western North Atlantic (Table 1) which lies close to the path of the Deep Western Boundary Current (the southern extension of NADW) and, hence, should sample 'well mixed' NADW on its passage south. The crust is up to 120 mm thick, shows no phosphatization ($P \leq 0.4\%$; compare refs 12, 18) and little elemental variation, except for Al which shows a distinct break 58 mm from the outer surface (Fig. 1a). Crust sampling and isotopic analysis follows procedures reported elsewhere¹², and isotopic data are given in Table 1. $\Delta^{87}\text{Sr}$ ages (Fig. 1b), calculated using Sr isotope stratigraphy^{19–22}, indicate that the crust started to grow some 18 Myr ago, and preserves two distinct growth intervals. The first, from 18.2 ± 0.6 to 16.2 ± 0.6 Myr ago and the second, from around 10.2 ± 1.2 Myr ago to the present day. The marked break at 58 mm corresponds to a hiatus of some 6.0 Myr, and correlates with the shift in Al, relating either to cessation of growth or an interval of crust erosion. Crust precipitation is considered to occur when oxygen-rich deep water is brought into contact with Mn²⁺-rich water from the oxygen minimum zone. It is interesting to note that the two phases of crust growth correspond to periods when oxygen-rich NADW¹⁴ is considered to have been actively formed⁸. The outer surface does not yield any useful age information, giving an $^{87}\text{Sr}/^{86}\text{Sr}$ ratio much more radiogenic than the present-day value for sea water, possibly due to the incorporation of detrital phases, or surface contamination during sampling. However, a $^{234}\text{U}/^{238}\text{U}$ age of 8 ± 0.3 kyr for the same sample²³ indicates that there was no significant break in growth, or loss of material. The sample contains no detectable secondary phases, or evidence for diagenesis, although it is not possible to rule out the possibility of isotopic homogenization after growth. However, the preservation of a major growth hiatus that correlates with the break in Al concentration, taken with the systematic nature of the $\Delta^{87}\text{Sr}$ ages, suggests that there has been no significant perturbation of the elemental record. Thus, the ages can be

[†] Present address: Laboratoire de Géochimie et Cosmochimie, CNRS URA 1758, IPG-Paris, 4 Place Jussieu, 75252 Paris Cedex 05, France.

considered to constrain the variations of Nd and Pb in sea water over the past 18 Myr.

The Nd isotope data (expressed as ϵ_{Nd}) are plotted against $\Delta^{87}\text{Sr}$ age in Fig. 2a. The outer surface of the crust yields an ϵ_{Nd} value of -12.93 ± 0.20 , indistinguishable from present-day sea water in this area¹⁵. Between the present-day and a depth corresponding to 3–4 Myr the Nd shows a marked shift towards more radiogenic values, with no measurable change over the remaining time interval

of growth. Pb isotope ratios show a similar pattern (Fig. 2b) where the outer surface has a radiogenic Pb isotope composition shifting to a less radiogenic value some 3–4 Myr ago, thereafter Pb isotope ratios remain constant. Nd isotope records have recently been obtained from the central Pacific Ocean¹², and a typical profile is shown in Fig. 2a. From 20 Myr to around 3–4 Myr the sample shows an overall trend towards more radiogenic Nd isotope compositions, but around 3–4 Myr there is a clear reversal towards unradiogenic

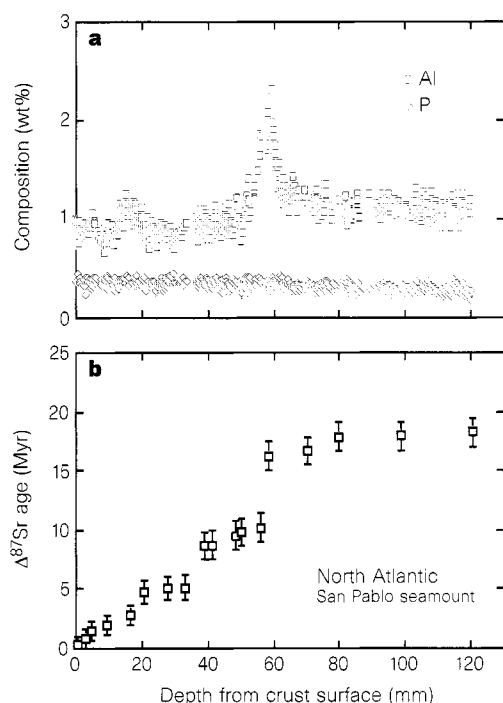


Figure 1 a, Chemical profiles of aluminium (squares) and phosphorous (diamonds) against depth for Fe-Mn crust BM1969.05 (see Table 1 for details of crust). In general, the crust shows little elemental variation, except for Al which shows a marked break ~58 mm from the outer surface, shifting from mean values of 1.1 wt% to values of 0.9 wt% for the outer 58 mm. The crust shows no evidence of secondary phosphatization, with $P \leq 0.4$ wt% for all intervals of growth (compare refs 12, 18). b, $\Delta^{87}\text{Sr}$ ages (Myr) against depth (mm) for sample BM1969.05. These ages were calculated using Sr isotope stratigraphy (data from refs 19–22) and indicate that the crust started to grow around 18.2 Myr ago, and preserves two distinct growth intervals. The first, from ~18.2 to 16.2 Myr, during which time the crust grew at an average rate of 30 mm Myr^{-1} , and the second, from 10.2 Myr to the present day, during which the crust grew at an average rate of 5.7 mm Myr^{-1} . The preservation of a major growth hiatus, of some 6.0 Myr, that correlates with the shift in Al argues against the possibility of isotopic homogenization.

Table 1 Isotope data for Fe-Mn crust BM1969.05

Sample No.	Depth in crust (mm)	$^{87}\text{Sr}/^{86}\text{Sr}^*$	$^{143}\text{Nd}/^{144}\text{Nd}^*$	$^{206}\text{Pb}/^{204}\text{Pb}$	$^{207}\text{Pb}/^{204}\text{Pb}$	$^{208}\text{Pb}/^{204}\text{Pb}^\dagger$	$\Delta^{87}\text{Sr}$ age, Myr ‡
1	0–1	0.709694 ± 20	0.511975	19.219	15.630	39.259	0.251
2	2–4	0.709130 ± 12	0.512005	18.976	15.609	38.828	0.82 ± 0.6
3	4–5	0.709098 ± 18	0.512038	18.966	15.637	38.883	1.45 ± 0.6
4	9–10	0.709074 ± 8	0.512047	18.898	15.598	38.765	1.92 ± 0.6
5	16–17	0.709032 ± 12	0.512063	18.880	15.623	38.816	2.74 ± 0.6
6	20–21	0.709012 ± 8	0.512078	18.841	15.587	38.701	4.77 ± 0.5
7	27–28	0.708969 ± 8	0.512078	18.825	15.580	38.657	5.09 ± 0.5
8	32–34	0.708965 ± 10	0.512083	18.811	15.614	38.774	5.13 ± 0.5
9	38–40	0.708891 ± 8	0.512092	18.798	15.582	38.658	8.64 ± 1.2
10	40–43	0.708889 ± 8	0.512089	18.822	15.599	38.714	8.70 ± 1.2
11	48–49	0.708861 ± 10	0.512098	18.805	15.587	38.670	9.55 ± 1.2
12	50–51	0.708852 ± 10	0.512080	18.832	15.622	38.788	9.82 ± 1.2
13	55–57	0.708839 ± 8	0.512072	18.758	15.580	38.634	10.21 ± 1.2
14	58–59	0.708718 ± 8	0.512084	18.770	15.598	38.701	16.23 ± 0.6
15	70–71	0.708679 ± 8	0.512094	18.765	15.571	38.623	16.66 ± 0.6
16	79–81	0.708673 ± 10	0.512080	18.809	15.596	38.697	17.84 ± 0.6
17	98–99	0.708567 ± 12	0.512072	18.822	15.617	38.763	17.91 ± 0.6
18	120–121	0.708537 ± 8	0.512082	18.787	15.571	38.604	18.24 ± 0.6

The crust analysed here was taken from San Pablo seamount, in the northwestern Atlantic Ocean ($39^\circ 0' \text{ N}$, $60^\circ 57' \text{ W}$; water depth ~1,850 m).

Quoted errors for the $^{87}\text{Sr}/^{86}\text{Sr}$ ratio are $2\sigma_{\text{r}}$. All $^{143}\text{Nd}/^{144}\text{Nd}$ and Pb isotope ratios gave an internal $2\sigma_{\text{r}}$ error equal to or better than the external reproducibility of the respective standard analysis given below.

* $^{87}\text{Sr}/^{86}\text{Sr}$ normalized to $^{86}\text{Sr}/^{86}\text{Sr} = 0.1194$; NBS-987 standard yields 0.710235 ± 18 (2σ , external; $n = 28$).

† $^{143}\text{Nd}/^{144}\text{Nd}$ normalized to $^{146}\text{Sm}/^{144}\text{Nd} = 0.7219$; JN-1 standard yields 0.511122 ± 9 (2σ , external; $n = 28$), La Jolla Nd standard yields 0.511858 ± 10 (2σ , external; $n = 8$).

‡ Pb isotope ratios relative to NBS-981 standard, which yields $^{206}\text{Pb}/^{204}\text{Pb} = 16.932 \pm 0.012$, $^{207}\text{Pb}/^{204}\text{Pb} = 15.486 \pm 0.016$, $^{208}\text{Pb}/^{204}\text{Pb} = 36.691 \pm 0.036$ (2σ) ($n = 40$).

§ $\Delta^{87}\text{Sr}$ ages calculated using data from refs 19–22. The quoted age uncertainties are governed by (1) the rate of increase in $^{87}\text{Sr}/^{86}\text{Sr}$ for a given time interval, and the fit of the reference data for each given time interval; (2) The precision of the $^{87}\text{Sr}/^{86}\text{Sr}$ measurement, typically 20×10^{-6} for the reference data^{19, 22}, and 18×10^{-6} for the measurements here (see standard data above). Thus, the differences in quoted uncertainties are largely due to variations in the slope and uncertainty of the seawater Sr reference curve. These uncertainties take no account of the accuracy of the correlation between the Sr seawater curve and the geomagnetic polarity timescale, or sample integrity (see discussion in ref. 22).

¶ Interpolated age based on the average growth rate of 5.7 mm Myr^{-1} for the outer 57 mm of the crust, by comparison a $^{234}\text{U}/^{238}\text{U}$ age of $8 \pm 0.3 \text{ kyr}$ has been obtained for the same sample²³.

values. The timing of this change is indistinguishable to that observed in the North Atlantic sample studied here, and suggests a global response in the oceans to this unradiogenic Nd signal. In contrast, Pb isotope data for the same sample¹² shows little variation over the past 20 Myr (Fig. 2b).

There are four major sources of NADW today; Denmark Strait Overflow Water (DSOW) and Iceland–Scotland-ridge Overflow Water (ISOW) (the nordic sources) both produced by rapid downwelling of cold, saline, surface water; Labrador Sea Water (LSW) (the boreal source) a warmer, shallow, water mass, the lower layers of which mix by entrainment with the underlying DSOW and ISOW

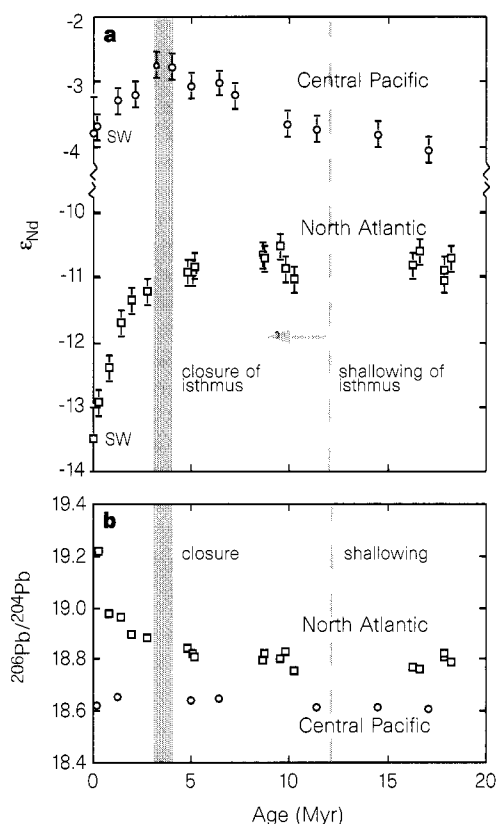


Figure 2 a, Nd composition (expressed as ϵ_{Nd}) against age (Myr) for the North Atlantic Fe–Mn crust (BM1969.05) studied here, and a crust from the central Pacific Ocean (CD29-2)¹². For the North Atlantic sample (squares) the outer surface yields an ϵ_{Nd} value indistinguishable from modern sea water (marked SW) at similar water depths¹⁵, typical of the value for ‘well mixed’ NADW¹⁵. The dramatic change over the past 3–4 Myr indicates a progressive enhancement of the isotopic signal associated with NADW, in particular water sourced in the Labrador Sea (see text). The Pacific sample (circles) grew over the same time interval as the Atlantic crust, and the outer growth surface also yields an ϵ_{Nd} value similar to present-day sea water³⁰ (marked SW). From 20 to 3–4 Myr the record shows an overall trend towards a more radiogenic Nd isotope composition, around 3–4 Myr ago there was a marked reversal of this trend towards unradiogenic Nd compositions (see text). $\epsilon_{Nd} = [(^{143}Nd/^{144}Nd)_{sample}/(^{143}Nd/^{144}Nd)_{CHUR} - 1] \times 10^4$, where $(^{143}Nd/^{144}Nd)_{CHUR} = 0.512638$, the ratio of the meteoritic chondrite uniform reservoir¹⁶ normalized to a $^{146}Nd/^{144}Nd$ ratio of 0.7219. **b**, $^{206}Pb/^{204}Pb$ ratios against age (Myr) for the same samples shown in **a**. For the Atlantic crust (BM1969.05) the outer surface yields a relatively radiogenic Pb isotope composition. This Pb signature, in particular the high $^{207}Pb/^{204}Pb$ ratios (Table 1) is characteristic of old continental crust (compare ref. 27), consistent with the unradiogenic Nd compositions for the same samples. Likewise, Pb shows a marked shift over the past 3–4 Myr, suggesting an increasing contribution from this old crustal source towards the present day. The Pacific sample (CD29-2) shows no measurable change, at this scale of sampling¹², over the past 20 Myr (see text).

water; and northward-moving Antarctic Bottom Water (AABW) which warms, shoals and mixes to contribute to NADW (see ref. 13 and references therein). Of these sources only LSW possesses a Nd isotope composition sufficiently unradiogenic ($\epsilon_{Nd} < -18$)^{15,24} to yield the signature of ‘well mixed’ NADW which has an ϵ_{Nd} value of -13.5 ± 0.5 ¹⁵ (other ϵ_{Nd} values are as follows: DSOW, -8.6 ; ISOW, -7.7 ; AABW, -8.5 ; refs 15, 17). The unradiogenic Nd in LSW is ultimately derived from Baffin Bay water ($\epsilon_{Nd} \approx -20$)²⁴ which flows into the Labrador Sea and is sourced in the erosion of old continental material that constitutes the surrounding land masses. At the present-day Labrador Sea water is the only known source capable of imposing an unradiogenic Nd isotopic composition on NADW, and simple calculations suggest that this water mass provides more than 30% of the Nd in modern NADW²⁴.

The dramatic shift in the Nd isotope composition of sea water documented here therefore indicates a progressive enhancement of the isotopic signal associated with NADW over the past 3–4 Myr. The timing of this change is remarkably similar to that observed in samples¹² from the central Pacific Ocean, and this is the first time over the past 20 Myr where both Atlantic and Pacific sea water respond in the same manner to the input of Nd. If this Nd was ultimately derived from the North Atlantic, then it must have travelled via the Antarctic Circumpolar Current. This not only requires an intensification of NADW flow, but also strongly suggests that the present global thermohaline circulation pattern was not established until this time. For the central Pacific record the earlier shift towards radiogenic Nd compositions around 12 Myr may relate to initial shallowing of the Central American Isthmus, which prevented the access of deep waters with unradiogenic Nd from the North Atlantic. Benthic foraminifera indicate²⁵ the presence of a sill at shallow water depths by 12 Myr, and shoaling of the carbonate compensation depth in the eastern Pacific 10 Myr ago has also been attributed to a reduction in deep-water exchange across the Panama gateway²⁶.

The radiogenic Pb composition for present-day sea water in the North Atlantic, in particular the high $^{207}Pb/^{204}Pb$ ratios, is also characteristic of old continental crust (compare ref. 27), in accord with the unradiogenic Nd compositions. Similarly, the shift in Pb isotopes over the past 3–4 Myr suggests an increasing contribution from this old crustal source. However, the origin of less radiogenic Pb sampled before 3–4 Myr remains uncertain. It could be due to an increasing contribution of Pb derived from the Pacific, (via the Panama gateway), a local hydrothermal source, or simply the signature of NADW with a reduced contribution from LSW. At present there are no means to distinguish between these possibilities, but in any case the data suggests that the contribution from North Atlantic sources was reduced before 3–4 Myr, entirely consistent with the Nd data. The notable absence of any change in the Pacific Pb record around 3–4 Myr ago is in accord with the short residence time for this element, $\sim 10^2$ yr, compared to that of Nd, $\sim 10^3$ yr. As a consequence the Nd signal of NADW survives transport via the Antarctic Circumpolar Current, whereas the Pb signal is lost.

Sediment accumulation patterns⁷, carbon isotope data^{8,9} and Cd/Ca records^{10,11} all point to a general increase in NADW flow around 3–4 Myr ago. However, drift deposits in the North Atlantic result from deep water originating solely in the Nordic seas, and, at present, chemical records lack sufficient resolution to distinguish between water originating in the Boreal or Nordic seas. A simple threefold increase in the flux of both Boreal and Nordic sources will produce the required shift in ϵ_{Nd} composition. In this case, the isotopic record suggests gradually increasing NADW flow over the past 3–4 Myr, with maximum values for the present day. However, sediment patterns⁷ and carbon isotope gradients⁹ indicate that NADW flow 3 Myr ago was as strong or stronger than today. In this case, given the residence times for Nd and Pb, a rapid response in the isotopic record would be expected, rather than the gradual

change observed. Taken together, these data suggest that NADW flow attained levels similar to today 3–4 Myr ago, with a gradually increasing contribution from LSW, in particular after 2 Myr. Palynological evidence suggests that the near-surface Labrador current was not established until 1.8 Myr, consistent with increased convection at that time²⁸.

This study demonstrates the utility of long-lived radiogenic isotopes as tracers of water mass movement. Our results indicate that the isotopic signal associated with NADW strengthened around 3–4 Myr ago, with a gradually increasing contribution from water sourced in the Labrador Sea, and suggest that the present global thermohaline circulation pattern was not established until that time. These data cannot distinguish between increased erosional input into the Labrador Sea or increased deep-water formation in the same area. However, recent model simulations²⁹ provide support for an increased rate of deep-water formation in the Labrador Sea following closure of the Central American Isthmus. This result, when taken with our findings, supports the conclusion that isthmus closure had a major effect on circulation and climate. □

Received 25 October 1996; accepted 3 February 1997.

- Hall, M. M. & Bryden, H. L. Direct estimates and mechanisms of ocean heat transport. *Deep-Sea Res.* **29**, 339–359 (1982).
- Rintoul, S. R. South Atlantic interbasin exchange. *J. Geophys. Res.* **96**, 2675–2692 (1991).
- Boyle, E. A. The role of vertical chemical fractionation in controlling late Quaternary atmospheric carbon dioxide. *J. Geophys. Res.* **93**, 15701–15714 (1988).
- Curry, W. B., Duplessy, J.-C., Labeyrie, L. D. & Shackleton, N. J. Changes in the distribution of $\delta^{13}\text{C}$ of deep water CO_2 between the last glacial and the Holocene. *Paleoceanography* **3**, 317–341 (1988).
- Maier-Reimer, E., Mikolajewicz, U. & Crowley, T. J. Ocean General Circulation model sensitivity experiment with an open Central American Isthmus. *Paleoceanography* **5**, 349–366 (1990).
- Mikolajewicz, U., Maier-Reimer, E., Crowley, T. J. & Kim, K.-Y. Effect of Drake and Panamanian gateways on the circulation of an ocean model. *Paleoceanography* **8**, 409–426 (1993).
- Wold, C. N. Cenozoic sediment accumulation on drifts in the northern North Atlantic. *Paleoceanography* **9**, 917–941 (1994).
- Miller, K. G. & Fairbanks, R. G. In *The Carbon Cycle and Atmospheric CO_2 : Natural Variations Archaean to Present* (eds Sunquist, E. T. & Broecker, W. S.) 469–486 (Geophys. Monogr. Ser. 32, Am. Geophys. Union, Washington DC, 1985).
- Raymo, M. E., Hodell, D. & Jansen, E. Response of deep ocean circulation to initiation of northern hemisphere glaciation (3–2 Ma). *Paleoceanography* **7**, 645–672 (1992).
- Boyle, E. A. Cadmium: chemical tracer of deep-water paleoceanography. *Paleoceanography* **3**, 471–489 (1988).
- Delaney, M. L. Miocene benthic foraminiferal Cd/Ca records: South Atlantic and western equatorial Pacific. *Paleoceanography* **5**, 743–760 (1990).
- Ling, H.-F. et al. Evolution of Nd and Pb isotopes in Central Pacific seawater from ferromanganese crusts. *Earth Planet. Sci. Lett.* **146**, 1–12 (1997).
- Dickson, R. R. & Brown, J. The production of North Atlantic Deep Water: sources, rates, and pathways. *J. Geophys. Res.* **99**, 12319–12341 (1994).
- Broecker, W. S., Blanton, S., Smethie, W. M. & Ostlund, G. Radiocarbon decay and oxygen utilization in the Deep Atlantic Ocean. *Glob. Biogeochem. Cycles* **5**, 87–117 (1991).
- Piepgas, D. J. & Wasserburg, G. J. Rare earth element transport in the western North Atlantic inferred from Nd isotopic observations. *Geochim. Cosmochim. Acta* **51**, 1257–1271 (1987).
- Jacobsen, S. B. & Wasserburg, G. J. Sm–Nd isotopic evolution of chondrites. *Earth Planet. Sci. Lett.* **50**, 139–155 (1980).
- Jeandel, C. Concentration and isotopic composition of Nd in the South Atlantic Ocean. *Earth Planet. Sci. Lett.* **117**, 581–591 (1993).
- Bau, M., Koschinsky, A., Dulski, P. & Hein, J. R. Comparison of the partitioning behaviours of yttrium, rare earth elements, and titanium between hydrothermal marine ferromanganese crusts and seawater. *Geochim. Cosmochim. Acta* **60**, 1709–1725 (1996).
- Hodell, D. A., Mueller, P. A., McKenzie, J. A. & Mead, G. A. Strontium isotope stratigraphy and geochemistry of the late Neogene ocean. *Earth Planet. Sci. Lett.* **92**, 165–178 (1989).
- Hodell, D. A., Mead, G. A. & Mueller, P. A. Variation in the strontium isotopic composition of seawater (8 Ma to present): Implications for chemical weathering rates and dissolved fluxes to the oceans. *Chem. Geol.* **80**, 291–307 (1990).
- Hodell, D. A., Mueller, P. A. & Garrido, J. R. Variations in the strontium isotopic composition of seawater during the Neogene. *Geology* **19**, 24–27 (1991).
- Hodell, D. A. & Woodruff, F. Variations in the strontium isotopic ratio of seawater during the Miocene: stratigraphic and geochemical implications. *Paleoceanography* **9**, 405–426 (1994).
- von Blanckenburg, E., O’Nions, R. K., Belshaw, N. S., Gibb, A. & Hein, J. R. Global distribution of beryllium isotopes in deep ocean water as derived from Fe–Mn crusts. *Earth Planet. Sci. Lett.* **141**, 213–226 (1996).
- Stordal, M. C. & Wasserburg, G. J. Neodymium isotopic study of Baffin Bay water: sources of REE from very old terranes. *Earth Planet. Sci. Lett.* **77**, 259–272 (1986).
- Duque-Caro, H. Neogene stratigraphy, paleoceanography and paleobiogeography in northwest South America and the evolution of the Panama seaway. *Paleogeogr. Paleoclimatol. Paleoecol.* **77**, 203–234 (1990).
- Lyle, M., Dadey, K. A. & Farrell, J. W. The Late Miocene (11–8 Ma) eastern Pacific carbonate crash: evidence for reorganization of deep-water circulation by the closure of the Panama gateway. *Proc. ODP Sci. Res.* **121**, 821–838 (1995).
- Abouchami, W. & Goldstein, S. L. A lead isotopic study of Circum-Antarctic manganese nodules. *Geochim. Cosmochim. Acta* **59**, 1809–1820 (1995).
- Mudie, P. J., De Vernal, A. & Head, M. J. *Geological History of the Polar Oceans: Arctic versus Antarctic* (eds Bleil, U. & Thiede, J.) 609–646 (Kluwer Academic, Dordrecht, 1990).
- Mikolajewicz, U. & Crowley, T. J. Response of a coupled ocean energy balance model to restricted flow through the Central American Isthmus. *Paleoceanography* (in the press).
- Piepgas, D. J. & Jacobsen, S. B. The isotopic composition of neodymium in the North Pacific. *Geochim. Cosmochim. Acta* **52**, 1373–1381 (1988).

Acknowledgements. We thank H. Buckley, Natural History Museum, London for assistance and provision of the sample studied here; S. Reed for assistance with the microprobe analysis; B. Bourdon, G. Henderson, A. Cohen, and F. von Blanckenburg for discussions; and K.W.B. thanks C. J. Allegre for providing the time and support to finish this study. This work was supported by the NERC and the Royal Society.

Correspondence should be addressed to K.W.B. (e-mail: burton@ipgp.jussieu.fr).

Effect of Danube River dam on Black Sea biogeochemistry and ecosystem structure

Christoph Humborg*, Venugopalan Ittekkot†, Adriana Cociasu‡ & Bodo v. Bodungen*

* Baltic Sea Research Institute, Seestrasse 15, 18119 Rostock, Germany

† Institute of Biogeochemistry and Marine Chemistry, University of Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany

‡ Romanian Institute for Marine Research, Boulevard Mamaia 300, 8700 Constanta, Romania

Rivers contribute significantly to the pollution and eutrophication that have caused drastic changes to the ecosystem of the Black Sea^{1–3}. Although damming is known to affect riverborne nutrient loads, and thus riverine ecosystems, evidence for significant effects in open coastal waters is sparse^{4–6}. Here we present long-term data sets of water and nutrient discharge from the River Danube to the Black Sea. These data reveal a reduction in the dissolved silicate load of the river by about two-thirds since dam constructions in the early 1970s. A concomitant decrease in wintertime dissolved silicate concentrations by more than 60% was observed in central Black Sea surface waters. The consequent changes in silicon to nitrogen ratio of the Black Sea nutrient load appear to be larger than those caused by eutrophication alone, and seem to be responsible for dramatic shifts in phytoplankton species composition from diatoms (siliceous) to coccolithophores and flagellates (non-siliceous). Our results strongly suggest that the damming of the Danube has been instrumental in causing the observed changes in Black Sea surface waters^{3,7–9}, and that the large number of dams in operation around the world today could similarly affect the food web structure and biogeochemical cycling in coastal seas.

Human interventions have caused a worldwide increase in river inputs of nitrogen (N) and phosphorus (P) to the coastal seas by more than a factor of four, leading to considerable eutrophication^{10,11} and to an increase in the frequency of unusual and/or noxious algal blooms¹². Shifts from diatoms to non-siliceous phytoplankton in the receiving waters of the Mississippi and the Rhine have been attributed to a change in Si:N:P ratios^{4,12–14} caused by excess of N and P, that is, cultural eutrophication. The construction of dams in rivers can also cause considerable reductions in nutrient loads owing to the removal of these nutrients in reservoir sediments (the ‘artificial-lake effect’)⁵. Whereas this removal might be overcompensated by anthropogenic nitrogen and phosphorus inputs downstream of the reservoir, no such compensation has been observed for silicate, for example, in the Nile⁵.

The River Danube, which contributes ~70% of the river inputs into the Black Sea, was dammed in 1970–72 approximately 1,000 km upstream at the Yugoslavia/Romania border by the ‘Iron Gates’ causing significant changes in the Danube’s discharge pattern¹⁵. Annual maxima in Danube water discharge and silicate concentration in spring deviates from the general trend observed in pristine rivers (Fig. 1), where silicate concentrations show very little temporal and regional variations probably due to buffering processes involving silicium-containing minerals (clays)^{16,17}. Spectral

analysis (Fig. 2) of the Danube data shows that the annual frequency band revealed the highest variance densities of water discharge and silicate concentration; both variables are highly coherent ($k = 0.86$, see Table 1). Thus, the concentration of silicate and other nutrients can be expressed as a function of water discharge (Table 1). Furthermore, a phase displacement of about one month (~ 30 days) occurs, whereby silicate leads the water discharge.

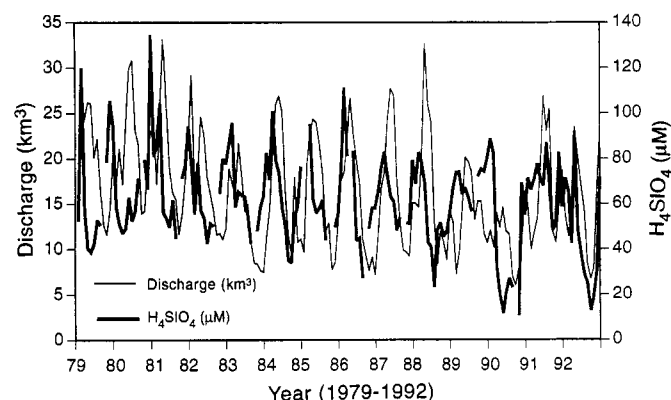


Figure 1 Silicate concentration (monthly means) and discharge (monthly means) of the River Danube from 1979 to 1992. The silicate time-series station (daily measurements) is located at the mouth of the Sulina branch of the Danube delta. Water discharge was measured ~ 60 km upstream of the river mouth, where the Danube enters the delta. Silicate samples were filtered through polycarbonate filters ($0.2 \mu\text{m}$) and frozen in 100-ml polyethylene bottles before being analysed by colorimetry¹⁹.

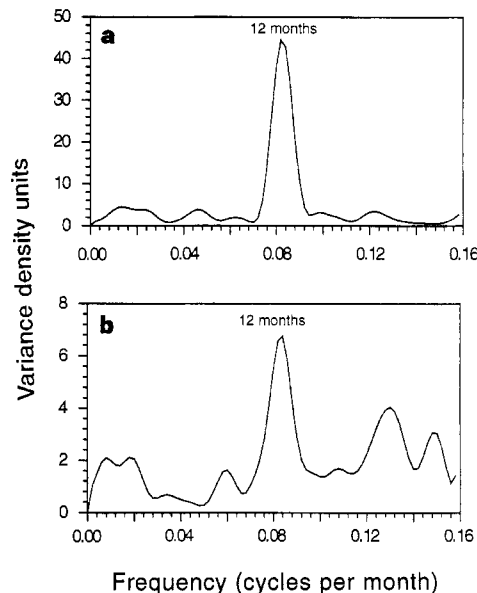


Figure 2 Variance density spectra of discharge (a) and silicate (b) time series 1979-92. Silicate data (monthly means) and water-discharge data (monthly means) were studied by time-series analysis to evaluate the statistical relationships between those time dependent variables. Spectral and cross-spectral analyses (Blackman-Turkey technique as used by Hays *et al.*²⁷; records filtered by cyclonic convolution procedures according to Mesko²⁸) of various time series were performed. Spectral analysis transforms the time series into a spectrum of dominant frequencies, where the variance spectrum represents the contribution of the whole variance spectrum over a specific frequency. It is therefore a relative measure of the amplitude of the dominant frequency of each specific time series. Main periodicities are indicated. Cross-spectral analysis for other nutrient variables is given in Table 1.

These two effects can be attributed to the pattern of reservoir operation. To prevent spring floods, the water stored in the reservoir is released through the outlets of the dam before maximum water levels are reached. Thus, nutrient-enriched reservoir waters flow into the Black Sea shortly before the annual peak in water discharge.

As well as the temporal variability in silicate discharge, the retention of silicate in the reservoir has led to an overall decrease

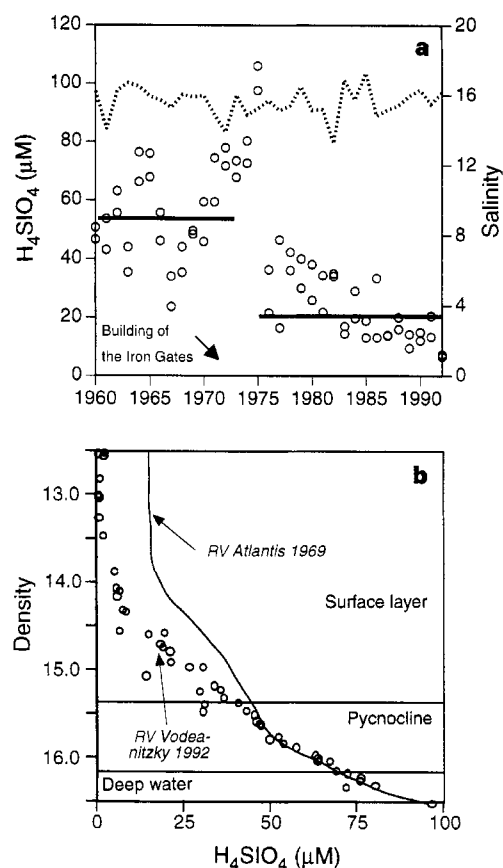


Figure 3 Silicate concentrations in the Black Sea. **a**, Mean winter silicate concentrations (open circles) at Constanta station (January and February, daily measurements); the bold lines gives the overall medians 1960-72 and 1973-92. Samples were measured directly after sampling using colorimetry^{29,30}. This time-series station is located about 60 nautical miles south of the Danube mouths. The salinity between 1960 and 1992 at Constanta station in winter (January and February, also daily measurements) has remained at about $S = 15$ (dashed line), indicating that the Danube influence on salinity has remained constant over time. **b**, Composite profiles of silicate plotted against density. Open circles, data from 7 stations of the 1992 RV Prof. Vodeanitzky (October); solid line, data from 1969 RV Atlantis II²⁴ (March-April). The surface layer in the central part of the Black Sea extends to the permanent pycnocline (a steep density gradient where density increases by $\sim 1\sigma_t$ within a few tens of metres) and corresponds to about 100 m water depth. Note that both the 1969 and 1992 data display conditions when dissolved inorganic nitrogen was depleted in the surface layer. Density was measured *in situ* with a Mark III CTD (conductivity, temperature and depth) instrument. Silicate was measured directly on-board by colorimetry²⁹.

in its concentration. The median silicate concentration for the period 1979–92 of 58 μM for the Danube (Fig. 1) is far less than the concentration of $\sim 140 \mu\text{M}$ derived from total discharge figures given by Almazov¹⁸ from measurements made during several cruises in 1959–60. These are the only data available before the construction of the 'Iron Gates' and show that the annual silicate load dropped from 800×10^3 tons (assuming Si is present as H_4SiO_4) in 1959–60¹⁸ to $(230\text{--}320) \times 10^3$ tons today¹. Winter concentrations of silicate in coastal waters covering the period 1960–92 decreased proportionally from 55 μM before the dam was built to 20 μM after its construction (Fig. 3a), which is in good agreement with the historical measurements given by Almazov¹⁸. Altered nutrient concentrations at this coastal station are clearly governed by the Danube river and in the case of nitrogen and phosphorus are only little influenced by coastal sources¹. The strong decrease in the Si:N ratio from 42 to about 2.8 can be attributed also to the concomitant rise in dissolved inorganic nitrogen caused by the increased Danube inputs (nitrate concentration at this coastal station increased from a median value of $\sim 1.3 \mu\text{M}$ in the early 1960s to a median of 7.9 μM in the 1980s¹).

These altered nutrient inputs have resulted¹⁹ in a distinct increase in phytoplankton bloom frequency, in cell densities and in the number of bloom-forming species from the beginning of the 1970s (Table 2). While diatom blooms increased by a factor of 2.5, blooms of non-diatoms such as dinoflagellates, the prymnesiophytes *Emiliania huxleyi* (coccolithophore) and the facultative toxic species *Chromulina* sp., as well as the Euglenophyte *Eutreptia lanowii*, increased by a factor of six. In our observations²⁰ in 1993, we found $140 \times 10^6 \text{ cells l}^{-1}$ of *Emiliania huxleyi* near the Danube plume at salinities as low as 10. Before the construction of the dams, blooms of coccolithophores were only reported from regions far offshore and later in the season²¹. The effect of the 'Iron Gates' on phytoplankton assemblages can be demonstrated by the following calculations: the annual increase of dissolved inorganic nitrogen load of the Danube since the early 1970s is estimated¹ to be $\sim 3.6 \times 10^{10}$ mol. Assuming Si:N ratios of 0.65 for *Skeletonema costatum*²², the predominant diatom species in the area (Table 2), a silicate uptake of $\sim 2.3 \times 10^{10}$ mol would be required to remove this nitrogen. Remarkably, this corresponds to the amount of silicate which is being retained in the reservoir each year (2.1×10^{10} mol). This implies that a large part of the past and future increase in nitrogen loads of the Danube after damming of the river are being (and will be) removed by non-diatom species.

The effect of silicate decrease is also seen in the central Black Sea, where the depth of the mixed layer varies both seasonally and regionally³, and diffusion and mixing along the isopycnal surfaces is rapid²³. We compare silicate concentrations from 1992 with histor-

ical data from the RV *Atlantis II* expedition in 1969^{9,24} by plotting silicate against density (Fig. 3b). Both were measured in periods of nitrogen depletion, when biological removal of silicate was at its limit. This comparison shows an order of magnitude decrease in silicate concentration, as was also reported in surface waters for 1988 (June–July) in the entire central Black Sea⁹.

The observed decrease above the pycnocline (corresponding to a depth of ~ 100 m) since 1969 amounts to $\sim 1,500 \text{ mmol m}^{-2}$ or 14.7×10^6 tons for the entire central basin. The total retention of silicate by the 'Iron Gates' of $\sim 11.8 \times 10^6$ tons since the construction of the dams in the early 1970s makes up 80% of this total decrease. Similar effects by dams on silicate discharge of the rivers are conceivable also for the Dnepr and Dnestr²⁵, which contribute to $\sim 20\%$ of river water input into the Black Sea. Thus, by far the largest part of the silicate reduction in the central basin appears to be caused by dam constructions around the Black Sea.

The long-range effect of the Danube is corroborated by data on barium (Ba) and radium (^{226}Ra) depletion in the Black Sea surface waters⁶; both these species are proxies for silicate. Model calculations⁸ implied that biological forcing, that is, sedimentation of diatoms, is primarily responsible for the decrease in Ba and ^{226}Ra (and thus for the decrease in silicate) in the central Black Sea over the past 30 years. However, satellite-derived chlorophyll data (coastal zone colour scanner data from 1,177 scenes, 1979–1982; EU Joint Research Centre) demonstrate that intense blooms (average chlorophyll concentrations, $3\text{--}10 \text{ mg m}^{-3}$) occur mostly in nearshore areas of the northwestern shelf, whereas in the central parts they do not exceed 0.5 mg m^{-3} . Moreover, Ba fluxes measured with sediment traps were lower by a factor of 10 than the model-predicted flux^{8,21}. The estimated decrease in Danube silicate inputs could account for this reported discrepancy.

Our results show that water and sediment storage in reservoirs behind the 'Iron Gates' have altered the biogeochemistry not just of the river and the adjacent coastal waters, but also of the entire Black Sea basin. The observed species shift towards carbonate-producing coccolithophores in the coastal waters (most certainly caused by decrease in silicate) exerts significant control on seawater chemistry such as alkalinity and pH. Furthermore, the occurrence of potentially

Table 1 Phase relationships of water discharge and nutrients

Variable	Phase displacement		Coherency	Zero-coherency	T
	ϕ (°)	months			
Discharge/ NO_2	26.7 ± 19.8	0.90	0.78	0.64	168
Discharge/ NH_4	26.1 ± 30.8	0.88	(0.60)	0.64	156
Discharge/ PO_4	22.7 ± 20.4	0.78	0.77	0.64	168
Discharge/ H_4SiO_4	29.8 ± 15.1	0.98	0.86	0.64	168

Phase relationships of cross-correlated variables (water discharge and nutrients) at the 12-month (annual) frequency band are given. Cross-correlation analyses the relationship of the different time-dependent variables; dominant frequencies of each variable and their phase relationships are studied. Phase displacement is described as the angle ϕ , which represents the contribution over a whole 360°-cycle. Coherency between two variables represents the correlation within the frequency domain, and is a common method of evaluating the statistical relationships of data sets within the time domain. Items tabulated are: phase displacement (ϕ) with 80% confidence interval; phase displacement in months coherency (k), entry in brackets indicates that cross-correlated variables are not coherent; 80% test statistics for non-zero coherency (k_0); and length of the time series (T) in months. Positive phases mean that the second variable leads the first variable. Cross-correlation of nitrate and water discharge was not carried out owing to the lack of sufficient numbers of nitrate measurements.

Table 2 Phytoplankton concentration in the northwestern Black Sea

	1960–70		1980–90	
	Cell densities ($10^6 \text{ cells l}^{-1}$)	Number of blooms	Cell densities ($10^6 \text{ cells l}^{-1}$)	Number of blooms
<i>Skeletonema costatum</i>	10–18	3	10–90	8
<i>Skeletonema subsalsum</i>	–	–	10–19	2
<i>Cyclotella caspia</i>	–	–	23–300	2
<i>Chaetoceros similis</i>	–	–	22	1
<i>Cerataulina pelagica</i>	–	–	5–6	3
<i>Nitzschia delicatissima</i>	6–21	4	17	1
<i>Nitzschia closterium</i>	–	–	13	1
<i>Nitzschia tenuirostris</i>	–	–	75	1
<i>Leptocylindrus danicus</i>	7	1	–	–
Total diatoms	7–21	8	5–300	19
<i>Prorocentrum cordatum</i>	17–51	4	10–810	9
<i>Prorocentrum scutellum</i>	–	–	7	1
<i>Scripsiella trochoidea</i>	–	–	26	1
<i>Heterocapsa triquetra</i>	–	–	5–12	3
Total dinoflagellates	17–51	4	5–810	14
<i>Eutreptia lanowii</i>	–	–	5–108	6
Total Euglenophytes	–	–	5–108	6
<i>Emiliania huxleyi</i>	–	–	220–300	2
<i>Chromulina</i> sp.	–	–	1,000	1
Total prymnesiophytes	–	–	220–1,000	3
Total blooms		12		42

Blooms are defined here as $> 5 \times 10^6 \text{ cells l}^{-1}$. Phytoplankton samples were taken¹⁹ daily at Constanta station, and monthly at offshore stations from different areas of the Romanian shelf. Cell counts were obtained by inverted microscopy. A dash indicates that the organism was absent.

toxic flagellate blooms may become more frequent, as observed in other coastal areas. Similar effects may be expected for the central Black Sea where, in contrast to pre-dam conditions, silicate depletion appears now to be more frequent. The combined effect of the altered silicate inputs from the Danube, overfishing and of the introduction of alien species such as ctenophores² on food-web structures is yet to be investigated. Our results further imply that other changes observed in the Black Sea, such as the spreading of the suboxic zone towards the surface⁹ or the increase in nitrate concentrations above the pycnocline⁷, may be related to alterations of the Danube input.

Today more than 36,000 dams are in operation around the world and more are being constructed at an appreciable rate²⁶. Though the observed effects of these dams can be expected to be more severe in closed seas (such as the Black Sea) than in rivers feeding into open systems, these effects could be of global consequence. □

Received 25 June 1996; accepted 5 February 1997.

- Cociasu, A., Dorogan, L., Humborg, C. & Popa, L. Long-term ecological changes in the Romanian coastal waters of the Black Sea. *Mar. Pollut. Bull.* **32**, 32–38 (1996).
- Mee, L. D. The Black Sea in a crisis: a need for concerted international action. *Ambio* **21**, 278–285 (1992).
- Murray, J. W. *et al.* Unexpected changes in the oxic/anoxic interface in the Black Sea. *Nature* **338**, 411–413 (1989).
- Admiraal, W., Breugem, P., Jacobs, D. M. L. H. A. & De Ruyter van Steveninck, E. D. Fixation of dissolved silicate and sedimentation of biogenic silicate in the lower river Rhine during diatom blooms. *Biogeochemistry* **9**, 175–185 (1990).
- Bennekou van, A. J. & Salomons, W. in *River Inputs to Ocean Systems* (eds Martin, J.-M., Burton, J. D. & Eisma, D.) 33–51 (UNEP, IOC, SCOR, United Nations, New York, 1981).
- Conley, J. D., Schelske, C. L. & Stroemer, E. F. Modification of the biogeochemical cycle of silica with eutrophication. *Mar. Ecol. Prog. Ser.* **101**, 179–193 (1993).
- Codispoti, L. A., Friederich, G. E., Murray, J. W. & Sakamoto, C. M. Chemical variability in the Black Sea: Implications of continuous vertical profiles that penetrated the oxic/anoxic interface. *Deep-Sea Res.* **38**, 691–710 (1991).
- Falkner, K. K., O'Neill, D. J., Todd, J. E., Moore, W. S. & Edmond, J. M. Depletion of barium and radium-226 in Black Sea waters over the past thirty years. *Nature* **350**, 491–494 (1991).
- Tugrul, S., Basturk, O., Saydam, C. & Yilmaz, A. Changes in the hydrochemistry of the Black Sea inferred from density profiles. *Nature* **359**, 137–139 (1992).
- Martin, J.-M., Burton, J. D. & Eisma, D. (eds) *River Inputs to Ocean Systems* (UNEP, IOC, SCOR, United Nations, New York, 1981).
- Nixon, S. W. Coastal eutrophication: a definition, social causes and future concerns. *Ophelia* **41**, 199–219 (1995).
- Smayda, T. J. in *Toxic Marine Phytoplankton* (eds Graneli, E., Sundström, B., Edler, L. & Anderson, D. M.) 29–41 (Elsevier Science, New York, 1990).
- Officer, C. B. & Ryther, J. H. The possible importance of silicon in the marine eutrophication. *Mar. Ecol. Prog. Ser.* **3**, 83–91 (1980).
- Turner, R. E. & Rabalais, N. N. Coastal eutrophication near the Mississippi river delta. *Nature* **368**, 619–621 (1994).
- Popa, A. Liquid and sediment inputs of the Danube River into the north-western Black Sea. *Mitt. Geol.-Paläontol. Inst. Univ. Hamburg* **74**, 137–149 (1993).
- Liss, P. S. in *Estuarine Chemistry* (eds Burton, J. D. & Liss, P. S.) 118–127 (Academic, New York, 1976).
- Edwards, A. M. C. & Liss, P. S. Evidence for buffering of dissolved silicon in fresh waters. *Nature* **243**, 341–342 (1973).
- Almazov, N. M. Stok rastverennykh soley i biogennykh veshchestv kotorye vynoseatsya rekami USSR. v Chernoe More. *Naukovi Zapiski Odes. Biol. St.*, Kiev **3**, 99–107 (1961).
- Bodeanu, N. Structure et dynamique de l'algoflore unicellulaire dans les eaux du littoral roumaine de la Mer Noire. *Cercetari Marine* **20/21**, 19–250 (1987–88).
- Humborg, C. Untersuchungen zum Verbleib der Nährstofffrachten der Donau. *Ber. Inst. Meereskunde, Kiel* **264**, (1995).
- Hay, B. J. Particle flux in the western Black Sea in the present and over the last 500 years: Temporal variability, sources, transport mechanisms. Thesis, Woods Hole Oceanographic Inst. (1987).
- Brzezinski, M. A. The Si:C:N ratio of marine diatoms: interspecific variability and effects of some environmental variables. *J. Phycol.* **21**, 347–357 (1985).
- Oguz, T. *et al.* Circulation in the surface and intermediate layers of the Black Sea. *Deep-Sea Res.* **40**, 1597–1612 (1993).
- Brewer, P. G. *Hydrographic and Chemical Data from the Black Sea* 75–66 (Tech. Note, Woods Hole Oceanogr. Inst., Woods Hole, 1971).
- Tolmazin, D. Changing coastal oceanography of the Black Sea. I: The northwestern shelf. *Prog. Oceanogr.* **15**, 217–276 (1985).
- Goudi, A. *The Human Impact on the Natural Environment* 177–234 (MIT Press, Cambridge, MA, 1994).
- Hays, J. D., Imbrie, J. & Shackleton, N. J. Variations in the earth's orbit: pace maker of the ice ages. *Science* **194**, 1121–1132 (1976).
- Mesko, A. *Digital Filtering Applications in Geophysical Exploration for Oil* (Wiley, New York, 1984).
- Koroleff, F. in *Methods of Seawater Analysis* (eds Grasshoff, K., Ehrhardt, M. & Kremling, K.) 125–187 (Verlag Chemie, Weinheim, 1983).
- Chirila, V. in *Ecologie Marina* Vol. 1 (ed. Bacescu, M.) 139–185 (Academiei Republicii Socialiste Romania, Bucuresti, 1965).

Acknowledgements. We thank U. Struck and S. Podewski for statistical analyses; J. Murray for comments on the manuscript; L. Popa and L. Dorogan for parts of the nutrient determinations; S. Kempe and A. Reimer for cruise coordination; and U. Horstmann for encouragement and support. This work was supported by the German Volkswagen Foundation and the EU EROS (European River Ocean Study) Black Sea Programme.

Correspondence should be addressed to V.I. (e-mail: ittekkot@dkrz.de).

Evidence for a selectively favourable reduction in the mutation rate of the X chromosome

Gilean T. McVean* & Laurence D. Hurst†

* Department of Genetics, Downing Street, Cambridge CB2 3EH, UK

† School of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

The equilibrium per-genome mutation rate in sexual species is thought to result from a trade-off between the benefits of reducing the deleterious mutation rate and the costs of increasing fidelity^{1,2}. We propose that selection will often favour a lower mutation rate on the X chromosome than on autosomes, owing to the exposure of deleterious recessive mutations on hemizygous chromosomes. We tested this hypothesis by examining 33 X-linked genes that have been sequenced in both mouse and rat, and compared their rate of evolution against 238 autosomal genes. The X-linked genes were found to have a significantly lower rate of synonymous substitution than the autosomal genes. Neither the supposed higher mutation rate in males nor stronger purifying selection against slightly deleterious mutations on the X chromosome can account for the low value. The most parsimonious explanation is that rodents have a lower mutation rate on the X chromosome than on autosomes. It is therefore likely that previous indirect estimates of the excess male mutation rate are inaccurate. Indeed, after correction we find no evidence for a male-biased mutation rate in rodents. Furthermore, the rate of synonymous substitution in Y-linked genes is not significantly different from that in autosomal ones. The extent to which enhanced male mutation rates are problematic³ for the mutational deterministic model⁴ of the evolution of sex must, in turn, be questioned.

The evolution of mutation rates is governed by two opposing forces: the costs imposed by deleterious mutations, and the ability to adapt to a changing environment^{1,5}. In contrast to asexual populations, selection acting within a sexual population will always favour a mutation rate of zero¹. That this is not observed can be accounted for by supposing that there is a trade-off between the benefits of reducing the deleterious mutation rate, and the costs imposed by increasing fidelity (such as the time and energy spent proof-reading)^{1,2}. Alternatively, there might be a physiological limit to the degree of accuracy in DNA replication, or perhaps group selection favours those lineages capable of producing new mutations and hence adapting.

The trade-off theory predicts that the selection acting on a modifier of mutation rates is proportional to the strength of selection against a mutant allele. Owing to the exposure of deleterious recessive mutations on hemizygous chromosomes, it follows that the selective advantage of an unlinked modifier reducing the mutation rate of the X chromosome can, under many circumstances, be much higher than that of a modifier reducing the autosomal mutation rate by the same factor (Fig. 1 and Box 1). This need not always be true. In mammals, more mutations are thought to occur in the male germ line than in the female⁶. As the proportional excess male mutation rate (α) increases, so the number of deaths imposed by new mutations on autosomes and on the X will tend to be comparable, as autosomes spend more time in males than does the X chromosome. Further, as the strength of selection on the modifier depends on the number of loci under its influence, a reduced mutation rate on the X chromosome is more likely in organisms with proportionally large X chromosomes and a

high mutation rate. As long as the proportional excess male mutation rate (α) is not too high, the X chromosome is adequately large, and the mutation rate high enough, we surmise that organisms should be adapted to ensure a lower mutation rate of X-linked genes than autosomal ones.

Comparison of homologous X- and Y-linked sequences indicates that rodents have a low proportional excess male mutation rate ($\alpha \sim 2$)^{7,8}. They also have a large X chromosome and a high mutation rate. Given, in addition, that many genes have been sequenced in both mice and rats, rodents are a good testing ground for the above theory.

From our analysis of genes on X chromosomes and autosomes (Fig. 2), the rate of substitution of synonymous mutations in X-linked genes (K_{sX}) is 14.63 ± 0.85 , much lower than that in autosomal ones (K_{sA}), 23.44 ± 0.44 (mean $\pm$ s.e. adjusted for sequence

length; all K values given are per 100 sites). The ratio of the rate of substitution of synonymous mutations in X-linked genes to that in autosomal ones is hence 0.62 ± 0.04 (s.e. calculated by the delta technique⁹), which is consistent with X-linked genes having a reduced mutation rate.

Two alternative hypotheses must be considered¹⁰. First, if synonymous mutations were not effectively neutral but slightly deleterious, then the low value of K_{sX} could be due to enhanced purifying selection¹⁰. However, if synonymous mutations have any deleterious effects these are likely to be very slight¹¹ and so additive¹². As a consequence, the relative intensity of selection against slightly deleterious mutations on the X chromosome and autosomes is not expected to be greatly different. More importantly, available evidence 'strongly suggests'¹³ that synonymous substitutions in rodents are effectively neutral¹³⁻¹⁶. For example, synonymous substitution rates are comparable to rates of divergence in pseudogenes¹⁵ and in the genome as a whole¹³ (which is largely non-coding). Similarly, selection does not seem to affect codon usage patterns in most mammalian genes^{13,16}, and the generation time effect is more conspicuous for synonymous than for non-synonymous substitutions^{17,18}. Further, we have compared synonymous substitution rates in all available coding and non-coding regions (intronic and 5' region) of the rodent Y chromosome, and found no difference between the two ($K_{sYcoding}/K_{Ynon-coding} = 1.02$). Assuming that substitutions in non-coding DNA are neutral, we conclude that synonymous substitutions in coding sequence are also neutral. Finally, if silent sites were slightly deleterious, then a

Box 1 Selection on modifiers of mutation rate.

Our conclusions follow from a simple extension of Leigh's model¹ for the evolution of mutation rates in which we examine the invasion of a costly modifying gene that reduces the mutation rate of a locus that is either autosomal or X-linked in a diploid sexual population. Given that the basic properties of Leigh's model¹ are robust to more realistic assumptions², we consider this extended model adequate for our purposes. The locus under mutation has two alleles, A and a (the latter is the deleterious, mutated form), the modifier locus freely recombines with A/a, and affects the germline mutation rate of the A allele. This modifier locus also has two alleles, M and m, with no mutation. The relative sex-specific viability of individuals is determined by their composition at the A/a and M/m loci: if AA or Aa (hemizygous) they have fitness 1; if aa or a then fitness $1-s$; and if Aa then fitness $1-hs$; if MM they have fitness $(1-c)^2$; if Mm then fitness $1-c$; and if mm then fitness 1; where c is the cost of mutation rate reduction. Fitness is multiplicative, so an individual that is AaMM has fitness $(1-hs)(1-c)^2$. Gametes are subject to sex-specific mutation rates, converting A to a, but not vice versa. In the absence of the modifier (in mm individuals), females have a mutation rate μ and males a rate $\alpha\mu$. The modifier is dominant and reduces these rates by a factor β : MM and Mm males have a mutation rate $\beta\alpha\mu$, and females a rate $\beta\mu$. We then ask whether, starting from mutation-selection equilibrium for the A/a locus in the absence of M, a rare M allele can invade.

Following the approach of Leigh¹, the selective advantage of a rare, cost-free modifier that reduces the mutation rate at an unlinked autosomal locus in a diploid sexual population at mutation-selection equilibrium approximates to $\mu(1-\beta)(1+\alpha)hs(1-\hat{q}_a)$, where $\hat{q}_a$ is the equilibrium frequency of the deleterious allele and the mean fitness of the population is $1-hs\hat{q}_a$. The selective advantage of the same modifier affecting an X-linked locus is $(2\mu/9)(1-\beta)(2+\alpha)s(1+2h)(1-\hat{q}_x)$, where $\hat{q}_x$ is the equilibrium frequency of the deleterious X-linked allele and the mean fitness of the population is $1-(s/3)(1+2h)\hat{q}_x$. The ratio of selective advantage (and hence maximum cost the modifier can suffer and still invade) for a modifier affecting an X-linked compared to an autosomal locus is thus approximately $\frac{2(2+\alpha)(1+2h)(1-\hat{q}_x)}{9(1+\alpha)h(1-\hat{q}_a)}$. Furthermore, given

that $\hat{q}_a \approx \frac{\mu(1+\alpha)}{2hs}$, and $\hat{q}_x \approx \frac{\mu(2+\alpha)}{s(1+2h)}$, it follows that this ratio can be expressed simply in terms of the excess male mutation rate and parameters of the locus under mutation. The exact form of the relationship is determined by simulation and shown in Fig. 1.

We have modelled the effect of heritable (germline) mutations only. In multicellular organisms a reduction of the spontaneous somatic mutation rate of the X chromosome will also be beneficial. Many insects, however, polyploidize their somatic cells, and so will not be under such strong selection pressure to reduce the mutation rate of the X chromosomes. Similarly, species with few somatic cell divisions are less likely to have a reduced X-chromosome mutation rate than those with numerous somatic divisions.

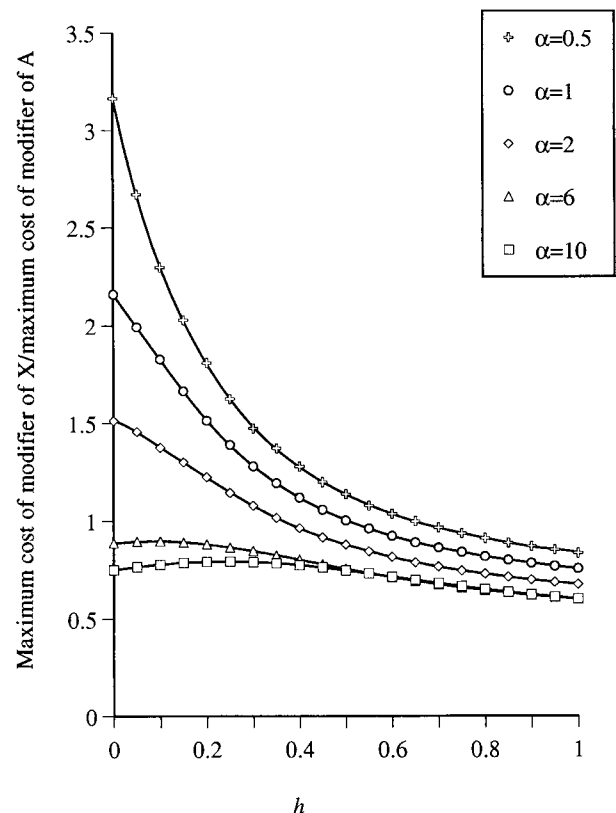


Figure 1 The ratio of the maximum cost that an autosomal modifier reducing the mutation rate of an X-linked gene can suffer and still invade, compared with that which the same modifier can suffer and affect an unlinked autosomal gene, for five values of the proportional male mutation rate (α) and for different values of deleterious mutation expressivity (h), where $s = 0.05$, $\mu = 0.001$ and $\beta = 0.5$. The form of this ratio is largely insensitive to the particular parameters used as long as $s > 0$, $\beta < 1$ and $\mu > 0$ (that is, deleterious mutations occur and the modifier reduces the mutation rate).

significantly higher rate of substitution would be expected in regions of low recombination with small population sizes (such as the Y chromosome), owing to the effects of background selection¹⁹ and other forces. A K_{SY} value of 24.63 ± 1.46 ($N = 6$) compares well with an autosomal rate of $K_{SA} = 23.44 \pm 0.44$. Thus the rate of synonymous substitution in Y-linked genes is not significantly higher than that in autosomal genes, consistent with their being effectively neutral (Fig. 3).

Furthermore, making no assumptions about the selective value of synonymous substitutions, we investigated whether purifying selection is the dominant force affecting substitutions on the X chromosome (other forces, such as reduced effective population size or low recombination rate, affect the rate of substitution of slightly deleterious genes^{19,20}). Given that selection acts more efficiently on non-synonymous than on synonymous mutations, it is expected that if purifying selection is the dominant factor, then, for a given K_{SX} , the value of K_{aX} would be predicted to be less than that of an autosomal gene with the same synonymous substitution rate. We find no evidence that this is so (Fig. 3). We compared the rank order in K_S of X-linked genes amongst autosomal ones with the rank order as regards K_a , and found that, for their K_S , X-linked genes have a higher K_a than would have been expected if the genes were autosomal (Mann-Whitney U test, $P < 0.05$, two-tailed). As the significance level is marginal, we conclude only that there is no evidence that purifying selection on the X-chromosome is the dominant force affecting the rate of substitution, and hence it is unlikely to explain the low value of K_{SX} .

Second, because the X spends less time in males than does an autosome, the reduction could result solely from the enhanced mutation rate in males⁶. We could then calculate α (the ratio of male to female mutation rates) from the ratio of K_S for X and autosomes given that $K_{SX}/K_{SA} = 2(\alpha + 2)/3(\alpha + 1)$. From this, we obtain the figure $\alpha > \infty$, which is neither logically viable nor consistent with previous estimates^{7,8}, by preferable methodologies¹⁰, that use X-Y homologous sequences. Our estimate is over six standard errors away from the estimate that $\alpha = 2$, and has 95% confidence intervals at $\alpha = \infty$ and 16.8 (using 95% confidence values for K_{SX}/K_{SA} (that is, $0.62 \pm (0.04 \times 1.96)$) substituted into the above

equation). These intervals are not close to overlapping with the confidence intervals from the X-Y comparisons^{7,8}, the highest of which is $\alpha = 3.9$. The estimate of $\alpha > \infty$ is the same as previous estimates by the same method^{6,10}.

The estimate of $\alpha \sim 2$ was made without considering the possibility that the X chromosome might have a reduced mutation rate. But if the X has a lower rate than the Y, independent of sex-specific effects, then the comparison of X- and Y-linked genes could give the misleading impression that there was an enhanced male mutation rate. The estimate of $\alpha = 2$ is hence an upper estimate, and we can reject the notion that an increased male mutation rate can explain the low rate of synonymous substitution of X-linked genes.

Further, we can then ask what value of α would be obtained if there were no excess male mutation rate and if all of the differences between X and autosome were due to a specially reduced rate on the X. Y-linked genes would then evolve at the same rate as autosomal genes. From this assumption we can calculate the expected artefactual value of α , from an X-Y comparison, in the absence of an excess male mutation rate. Intriguingly, this results in the value $\alpha = 2.3$, comparable to that found by previous X-Y comparisons. The apparent enhanced male mutation rate in rodents may therefore be artefactual. If this were the case, we would expect that Y-linked and autosomal genes should have the same synonymous substitution rates; we find this to be so ($K_{SY} = 24.6 \pm 1.46$; $K_{SA} = 23.44 \pm 0.44$). In contrast, if $\alpha = 2$ then the expected value of K_{SY} is $1.33 \times K_{SA} = 31.3$, which is 4.6 standard errors from the value observed (and hence is highly significantly different) (Fig. 3). If α is at the low end of previous estimates (say, 1.5), then the value of K_{SY} is 2.4 standard errors away from that expected. We can therefore find no evidence for an excess male mutation rate of the magnitude of that previously reported. We do not propose that there is no excess male mutation rate in every mammalian species, but suggest that previous calculations of the excess using X-Y comparisons may be overestimates.

The absence of evidence for a male excess runs contrary to the belief that the number of cell divisions before meiosis determines the mutation rate^{6,7,18}. However, evidence is accumulating that this parameter is not always the sole determinant of the mutation rate in

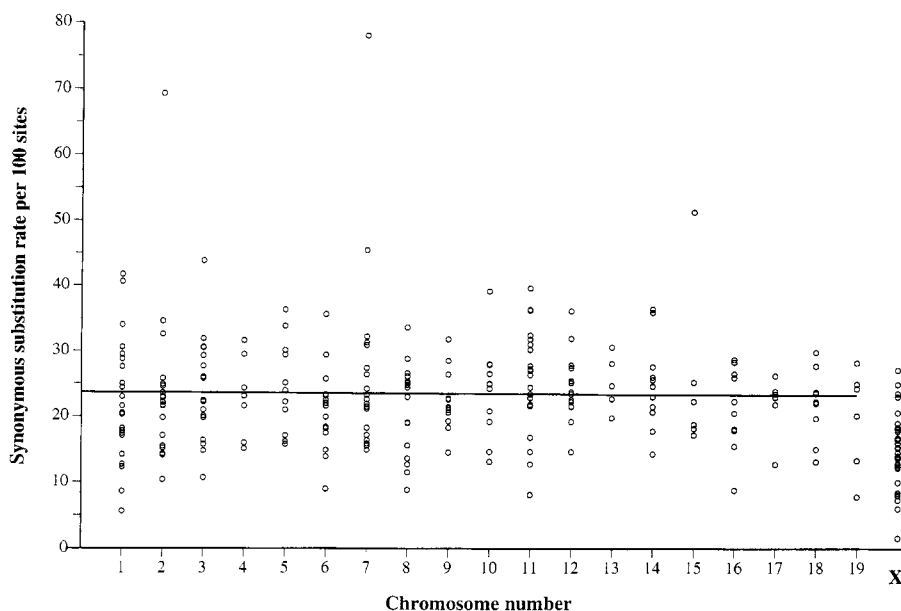


Figure 2 The values of the synonymous substitution rate per 100 sites (K_S) for 238 autosomal genes, arranged by chromosome, and 33 X-linked genes. Rodent chromosomes are arranged by size (the largest being chromosome 1 and the smallest chromosome 19). The horizontal line is the regression of chromosome

number (approximately proportional to the inverse of chromosome size) and K_S , showing that the lower rate of synonymous substitution of X-linked genes cannot be a function of chromosome size.

mammals, and that the origin of any sex differences in mutation rate is more complex. For example, the germline mutation rate of minisatellites (some of which show a male bias) is much higher than the somatic rate, so mutations in minisatellites are considered mostly meiotic in origin²¹. Similarly, reanalysis of X-linked haemophilia B has suggested that not all forms of mutation show a male excess²². For the forms of mutation that do show a male excess, the bias can be attributed to differences in germline methylation patterns rather than the number of cell divisions before meiosis²². This finding also suggests a possible mechanism for the reduced X-linked mutation rate, as both X chromosomes are unmethylated in female germ line²³. Perhaps ironically, if this were the case, X-linked mammalian genes should actually have a higher mutation rate in males than in females, as the X is methylated in male germ line. It follows that it is inappropriate to extrapolate from any sex bias in mutation rates at X-linked sites to any potential sex bias at other genomic locations.

Our finding has potential influence on theories of the evolution of sex. Enhanced male mutation rates are considered problematic³ for the mutational deterministic⁴ theory of sex, as under many circumstances females are better off being asexual and hence not having their progeny affected by large numbers of male-derived mutations. The extent of the problem needs re-evaluating (see also ref. 24).

The finding is also relevant to theories of speciation. One explanation of Haldane's rule (the inviability/sterility of the heterogametic sex in hybrids) holds that it is caused by rapid adaptive evolution of X-linked genes resulting from the exposure of advantageous recessives¹⁶. A reduced mutation rate of X-linked genes will lessen the difference in rates of adaptive evolution between X chromosome and autosome. Indeed, the absolute rate of substitution of non-synonymous mutations on the X is considerably lower than that on autosomes ($K_{aX}/K_{aA} = 0.434 \pm 0.13$) (Fig. 3). However, it is unclear what proportion of the X-linked substitutions are

due to enhanced selection for advantageous recessives. The finding that K_{aX} may be higher than expected after controlling for K_s (see Fig. 3) suggests that substitution of advantageous mutations may be common²⁵. □

Methods

Collection of sequence data. The X-linked genes were obtained from a list available on the mouse genome database²⁶ of nearly 200 confirmed X-linked (non-pseudoautosomal) genes on the mouse X chromosome. Orthologues were found by means comparable to previous methodologies¹³. Sequences were sought by searching the EMBL database by gene name through Sequence Retrieval System and by following leads in searches of the mouse genome database²⁶ by gene name. If orthologous pairs were obtained then sequences were extracted from EMBL. If both orthologues were not obtained from these searches, but a sequence was found in either mouse or rat, then this sequence was subjected to a BLAST search²⁷. Genes prescribed different names in mouse and rat were obtained by this method. Further, a few adequately similar (>90% identity) expressed sequence tags were obtained, although this biased the sample towards high homology. However, such genes are a limited part of our sample and, as our control sample is also biased in the same direction¹³, we do not consider this to be a problem. We found all of the 11 genes previously identified as orthologous¹³. The control sample was based on that assembled previously¹³, and linkage for each of the 363 genes in that sample was again ascertained through the mouse genome database²⁶. This resulted in the finding of 242 confirmed autosomal genes; four known imprinted genes were excluded from the autosomal set. Alignment of sequences was carried out using PILEUP from GCG²⁸. This was done both with nucleic acid sequence and protein sequence to ensure correspondence between the two. To enable direct comparison with previously described data, the extent of synonymous and non-synonymous divergence was calculated by the method of Li *et al.*²⁹ (the same as that used in ref. 13).

Comparison of synonymous sites and non-coding regions on the Y chromosome. The value of K_{sY} we obtained by analysis of coding sequence of rodent *Sry*, *Ube1y* and *Zfy* genes and one sample of non-coding sequence (intronic DNA or 5' region) from each gene. As there is no covariance between the non-coding and the coding components (adjusted $r^2 = 0.0$), we treat each observation as an independent estimate of the Y chromosome's synonymous substitution rate. The values for the non-coding sequence were found using the method of ref. 29 and corrected to render them comparable with estimates from coding sequence. The correction is necessary as application to coding sequence overestimates the frequency of synonymous changes at twofold degenerate sites. We converted the obtained data using the regression of the conversion data³⁰, such that if K_{sc} is the value of K_s comparable to those found using LWL on coding sequence, and K_{sf} is that obtained using LWL on non-coding sequence, then $K_{sc} = 1.078 K_{sf} + 0.0286$ (ref. 30). The value of K_{sY} is influenced heavily by one data point, namely the large intron of *Zfy*, and its removal adjusts the estimate to 22.7 ± 1.5 .

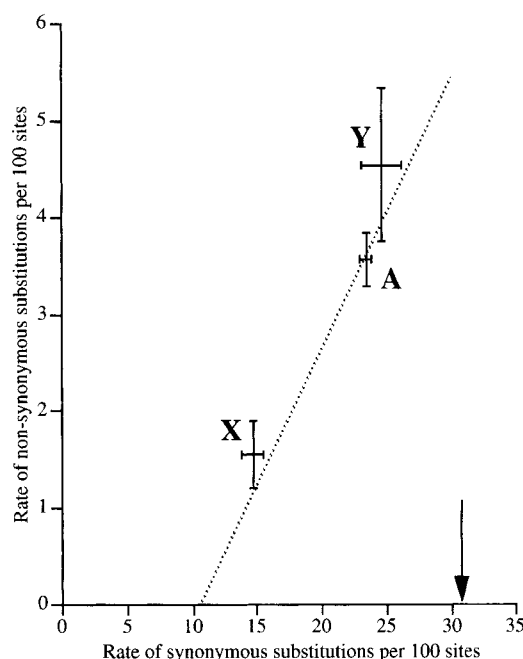


Figure 3 Rate of synonymous and non-synonymous substitutions per 100 sites for X-linked (X), Y-linked (Y) and autosomal (A) genes in the mouse/rat comparison (mean $\pm$ s.e.). The broken line represents the regression of K_s against K_a for autosomal genes. For their value of K_s , X-linked genes have a higher K_a than would be expected if they were autosomal (the regression line is only a rough guide to this, and the non-parametric statistic is reported in the text, $P < 0.05$). The arrow indicates the expected K_{sY} if $\alpha = 2$.

Received 11 September 1996; accepted 27 January 1997.

- Leigh, E. G. Natural selection and mutability. *Am. Nat.* **104**, 301–305 (1970).
- Kondrashov, A. S. Modifiers of mutation-selection balance: general approach and the evolution of mutation-rates. *Genet. Res.* **66**, 53–69 (1995).
- Redfield, R. J. Male mutation rates and the cost of sex for females. *Nature* **369**, 145–147 (1994).
- Kondrashov, A. S. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435–440 (1988).
- Kimura, M. On the evolutionary adjustment of mutation rates. *Genet. Res.* **9**, 23–34 (1967).
- Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. & Yasunaga, T. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 863–867 (1987).
- Chang, B. H. J., Shimmin, L. C., Shyue, S. K., Hewett-Emmett, D. & Li, W.-H. Weak male-driven molecular evolution in rodents. *Proc. Natl Acad. Sci. USA* **91**, 827–831 (1994).
- Chang, B. H. J. & Li, W.-H. Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked *Ube-1* genes and pseudogenes. *J. Mol. Evol.* **40**, 70–77 (1995).
- Bulmer, M. G. *Principles of Statistics* (Dover, New York, 1979).
- Shimmin, L. C., Chang, B. H. J., Hewett-Emmett, D. & Li, W.-H. Potential problems in estimating the male-to-female mutation-rate ratio from DNA-sequence data. *J. Mol. Evol.* **37**, 160–166 (1993).
- Li, W.-H. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345 (1987).
- Crow, J. E. & Simmonds, M. J. in *The Genetics and Biology of Drosophila* (eds Ashburner, M., Carson, H. L. & Thompson, J. N.) 1–35 (Academic, London, 1983).
- Wolfe, K. H. & Sharp, P. M. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37**, 441–456 (1993).
- Sharp, P. M. in *Evolution and Animal Breeding: Reviews on Molecular and Quantitative Approaches in Honour of Alan Robertson* (eds Hill, W. G. & Mackay, T. F. C.) 24–32 (CAB International, Wallingford, UK, 1989).

15. Wolfe, K. H., Sharp, P. M. & Li, W.-H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285 (1989).
16. Eyre-Walker, A. An analysis of codon usage in mammals: selection or mutation bias? *J. Mol. Evol.* **33**, 442–449 (1991).
17. Ohta, T. An examination of the generation time effect on molecular evolution. *Proc. Natl Acad. Sci. USA* **90**, 10676–10680 (1993).
18. Li, W.-H., Ellsworth, D. L., Krushkal, J., Chang, B. H. J. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation time effect hypothesis. *Mol. Phylog. Evol.* **5**, 182–187 (1996).
19. Charlesworth, B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63**, 213–227 (1994).
20. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130**, 113–146 (1987).
21. Jeffreys, A. J. et al. Complex gene conversion events in germline mutation at human minisatellites. *Nature Genet.* **6**, 136–145 (1994).
22. Ketterling, R. P. et al. Germ-line origins of mutation in families with hemophilia-B: the sex-ratio varies with the type of mutation. *Am. J. Hum. Genet.* **52**, 152–166 (1993).
23. Driscoll, D. J. & Migeon, B. R. Sex difference in methylation of single-copy genes in human meiotic germ-cells—implications for X-chromosome inactivation, parental imprinting, and origin of CpG mutations. *Somat. Cell Mol. Genet.* **16**, 267–282 (1990).
24. Hurst, L. D. & Peck, J. R. Recent advances in understanding of the evolution and maintenance of sex. *Trends Ecol. Evol.* **11**, A46–A52 (1996).
25. Gillespie, J. H. On Ohta's hypothesis: most amino acid substitutions are deleterious. *J. Mol. Evol.* **40**, 64–69 (1995).
26. Mouse Genome Database 3.1, Mouse Genome Informatics (Jackson Laboratory, Bar Harbor, ME, 1996). URL: <http://www.informatics.jax.org/>.
27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
28. GCG Wisconsin Package Version8, manual (Genetics Computer Group, Madison, WI, 1994).
29. Li, W.-H., Wu, C.-I. & Luo, C.-C. A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**, 150–174 (1985).
30. Li, W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**, 96–99 (1993).

Acknowledgements. We thank J. Barrett for statistical advice; B. Charlesworth, A. Kondrashov, W.-H. Li, M. Lyon, K. Wolfe and P. Sharp for comments on earlier versions of the manuscript; and M. Nei, C. Bishop and P. Keightley for assistance. L.D.H. is funded by The Royal Society, G.T.M. by the Medical Research Council.

Correspondence and requests for materials should be addressed to G.T.M. (e-mail: gatmv@mole.bio.cam.ac.uk).

Modular decomposition in visuomotor learning

Zoubin Ghahramani*† & Daniel M. Wolpert‡

* Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

† Present address: Department of Computer Science, University of Toronto, Toronto, Canada M5S 3H5.

‡ Sobell Department of Neurophysiology, Institute of Neurology, Queen Square, London WC1N 3BG, UK

The principle of 'divide-and-conquer', the decomposition of a complex task into simpler subtasks each learned by a separate module, has been proposed as a computational strategy during learning^{1–3}. We explore the possibility that the human motor system uses such a modular decomposition strategy to learn the visuomotor map, the relationship between visual inputs and motor outputs. Using a virtual reality system, subjects were exposed to opposite prism-like visuomotor remappings—discrepancies between actual and visually perceived hand locations—for movements starting from two distinct locations. Despite this conflicting pairing between visual and motor space, subjects learned the two starting-point-dependent visuomotor mappings and the generalization of this learning to intermediate starting locations demonstrated an interpolation of the two learned maps. This interpolation was a weighted average of the two learned visuomotor mappings, with the weighting sigmoidally dependent on starting location, a prediction made by a computational model of modular learning known as the 'mixture of experts'¹. These results provide evidence that the brain may employ a modular decomposition strategy during learning.

A general strategy for learning is to divide a complex task into simpler subtasks and learn each subtask with a separate module. This strategy has been formalized into a computational model of learning known as the mixture of experts¹, in which a set of expert

modules each learn one of the subtasks and a gating module weights the contribution of each expert module's output to the final system output. The gating module bases its weighting of each expert on its estimate of the probability that this expert is the appropriate one to use for the current task. During learning, the gating module simultaneously learns to partition the task into subtasks while the expert modules learn these subtasks. Such modular decomposition has been proposed both as a model of high-level vision⁴ and of the role of the basal ganglia during sensorimotor learning⁵. The mixture of experts model makes specific predictions regarding the nature of learning which have not been tested empirically. Here we test the hypothesis that the visuomotor system exhibits such modular decomposition during learning.

Previous studies have shown that the motor system is able to adapt to multiple different perturbations. Subjects adapt increasingly readily when presented repeatedly with two different prismatic displacements separated temporally^{6,7}, a process which is mediated by posterior parietal cortex⁸. Similarly, subjects adapt to multiple perturbations if cued by gaze direction^{9–11}, body orientation¹², arm configuration¹³, an auditory tone¹⁴ or the feel of prism goggles^{15–17}. One hypothesis to account for these studies is that multiple visuomotor mappings are stored simultaneously, suggesting a modular system. However, alternative explanations, such as a general increase in adaptability, or a single, non-modular system that is responsive to inputs from many modalities, cannot be ruled out from these studies. In particular, it is not clear whether the outputs of separate modules can be appropriately combined for contexts not already learned. Here we probe the existence of multiple modules by testing the specific predictions of a computational model of modular learning.

We investigated a learning paradigm in which the visual feedback of the hand was perturbed during pointing movements so that a

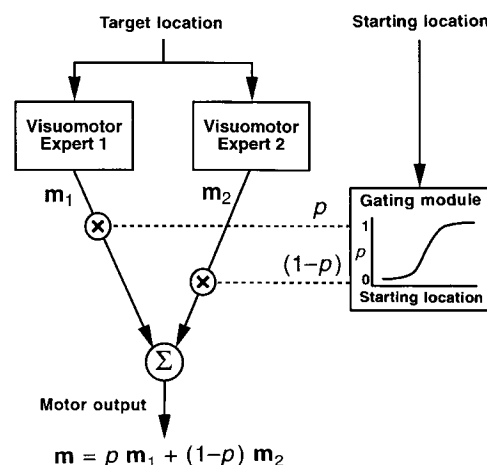


Figure 1 A modular decomposition model of visuomotor learning in which two different maps can be learned for the same visual target location. This represents the simplest instantiation of the hierarchical mixture of experts², having only one level and two experts. The model maps target and starting locations to motor outputs, $\mathbf{m}$, which could represent, for example, the final hand location or movement vector. Each expert learns a different mapping between target locations and motor outputs. The contribution of each expert's output, $\mathbf{m}_1$ and $\mathbf{m}_2$, to the final motor output, $\mathbf{m}$, is determined by the gating module's output, p . The output p reflects the probability that expert 1 is the correct module to use for a particular starting location; at p values of 1 or 0 the final output is determined solely by the output of expert 1 or expert 2 respectively, whereas at intermediate values of p both experts contribute to the final output. The logistic form of the gating module's output as a function of starting location can be derived by assuming that each expert learns the visuomotor map at one of the two starting locations—its preferred starting location—and that each expert is responsible for an equal size gaussian region around this preferred starting location.

single location in visual space was remapped to two different hand positions depending on the starting location of the movement (see Methods). This perturbation creates a conflict in the visuomotor map of the arm, the internal model¹⁸ of the kinematics of the arm which captures the normally one-to-one relation between visually perceived and actual hand locations^{19–21}. One way to resolve this conflict is to develop two separate visuomotor maps, the expert modules, each appropriate for one of the two starting locations (Fig. 1). A separate mechanism, the gating module, then combines, based on the starting location of the movement, the outputs of the two visuomotor maps. The output of the gating module, which represents the weighting given to each visuomotor map for a given starting location, has a sigmoidal (logistic) shape, as a function of the starting location of the movement (Fig. 1). This relationship results from the assumption that each expert is responsible for an equal variance gaussian region around its preferred starting location²², which corresponds to its receptive field. As in previous studies of the visuomotor system^{23–25}, the internal structure of the system can be probed by investigating the generalization properties in response to novel inputs, which in this case are the starting locations on which it has not been trained. The hallmark of a system with modular decomposition is the ability to learn both conflicting mappings, and to transition smoothly from one visuomotor map to the other in a sigmoidal fashion as the starting location is varied.

Subjects were exposed in a virtual reality setup to two different visuomotor perturbations, discrepancies between the actual and perceived hand location, depending from which of two possible starting locations the movement originated (L2 and L6 in Fig. 2; see Methods). Although subjects were unaware of the perturbation, they showed significant adaptive changes in their pointing behaviour when starting from locations L2 and L6 (Fig. 3b–d). The

adaptation seen for movements from these two points was significantly different from each other ($P < 0.001$), showing that the subjects were able to learn two distinct remappings of the same point in visual space as a function of the starting location. Furthermore, as the starting location was varied between L1 and L7, a smooth transition was seen in the change in pointing behaviour which reflects visuomotor learning (Fig. 3b–d).

We estimated the mixing proportion ('p' in Fig. 1) by fitting the changes in pointing behaviour at each starting location to a weighted mixture of the adaptation observed for movements starting from L6 and L2. These estimates show a significant modulation over the starting locations (Fig. 4) for groups b–d ($P < 0.001$), but as expected, not for the control group ($P > 0.05$), who showed no change in pointing behaviour. The modulation in groups b–d gave a significantly better fit to a logistic function, the mixing probabilities predicted by the modular decomposition model, than to a linear function ($P = 0.02$).

The hypothesis of modular decomposition can be contrasted with models in which a single visuomotor transformation is computed. Models in which the transformation is based solely on the visual location of the target cannot account for the two mappings learned for the same point in visual space. Alternatively, a single visuomotor transformation may take in as inputs both the visual location of the target and the movement starting location. The manner in which such a single module would generalize to new starting locations depends crucially on the internal structure of the module. For example, a linear constraint model²³ predicts a corresponding linear pattern of generalization, which was not observed in the data (Fig. 4b–d). In a previous study²⁵ we have shown that the visuomotor map shows limited generalization to a novel, starting-location-independent remapping, suggesting a local receptive field structure.

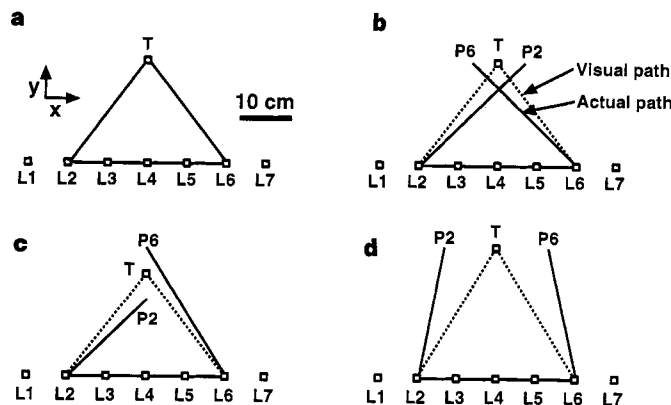


Figure 2 The perturbations used for the four groups of subjects (a–d). Movements were made in the horizontal plane and the schematic shows the seven possible starting locations (L1–L7) and the target (T) seen from above. Solid lines indicate the actual path taken by the hand during the exposure phase; dotted lines indicate the visually displayed path of the hand. For the control group (a) the two lines coincide everywhere as there was no perturbation and therefore no discrepancy between the visually displayed and actual hand location. For the perturbation groups, a discrepancy between displayed and actual hand position was introduced (see Methods for details). The discrepancy was chosen so that subjects, in order to perceive their hand visually on target T, had to point to two different locations, P2 and P6, depending on whether the movement started from L2 or L6.

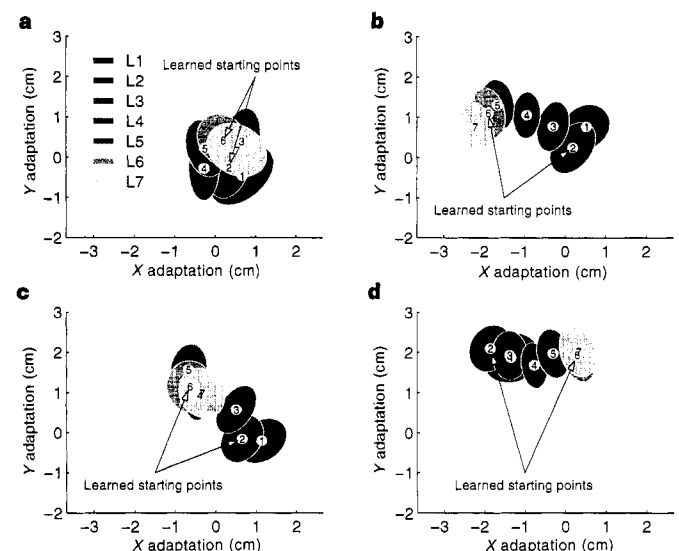


Figure 3 Adaptation of the visuomotor map for the control (a) and perturbation groups (b–d). For each starting location (L1–L7, denoted by shading), the 95% confidence ellipse for the change in pointing behaviour, induced by the visuomotor perturbation, is shown. For clarity, the centre of the ellipses, which represents the change in pointing behaviour, is also indicated numerically by the starting location (for example, 3 corresponds to the change in pointing for movements starting from L3). The change in pointing corresponding to the learned starting points L2 and L6 are indicated by the arrows. For the perturbation groups, significant changes in pointing are seen, corresponding to partial adaptation to the remappings introduced. These changes in pointing smoothly shift as the starting location varied between L1 and L7. As well as the changes in response to the perturbation, there was a starting-point-independent movement overshoot for the perturbation groups, accounting for both the Y offset of the means in b and d and the additional right-to-left shift seen in c.

If a single visuomotor module showed similarly local receptive fields, both in visual space and starting location, the predicted adaptation would be maximal at starting locations L2 and L6 and decay away from these points, a pattern not supported by the data. Our study indicates that two different maps can be learned for the same point in visual space and that the generalization to starting locations at which the subject was not exposed to the perturbation has the logistic relationship predicted by the mixture of experts model. These results provide evidence that modular decomposition is a feature of visuomotor learning.

These findings can be interpreted through the hypothesis that the visuomotor system maps visual vectors, pairs of target and starting locations, into movement vectors. Evidence for such vector-based coding has been obtained in neurophysiological studies which suggest that populations of cortical cells code for direction of movement^{26,27}. Similarly, it has been shown that a set of limb postures, which specify endpoints, can be achieved by stimulation of specific areas of the spinal cord, and that simultaneous stimulation of two such areas elicits a large repertoire of intermediate postures²⁸. According to either of these hypotheses, our results show that learning two new visuomotor mappings, whether represented as vectors or postures, at the two starting locations, leads to a smooth sigmoidal generalization at intermediate locations. This generalization is consistent with a gradual mixing, modulated by starting location, of two separate neuronal populations, each of which has learned a different visuomotor mapping. This suggests a simple and plausible neural mechanism by which the modular learning observed could have arisen in the visuomotor system. □

Methods

Thirty-two right-handed participants, who were naive to the purpose of the experiment and gave their informed consent, were randomly assigned to one of four groups: a, b, c and d.

Setup. Subjects sat at a large horizontal digitizing tablet with their head supported by a chin and forehead rest (a complete description of the set-up can be found in ref. 29). Subjects held a digitizing mouse with their right index finger tip mounted on its cross hairs — direct view of their arm was prevented

by a screen. The targets and feedback of hand position were presented as virtual images in the plane of the digitizing tablet, and therefore in the plane of the hand. This was achieved by projecting a computer display onto a horizontal rear projection screen suspended above the tablet. A horizontal front-reflecting mirror was placed face-up midway between the screen and the tablet. The subjects viewed the reflected image of the rear projection screen by looking down at the mirror. By matching the screen-mirror distance to the mirror-tablet distance, all projected images appeared to be in the plane of the hand when viewed in the mirror. Targets were represented as 1-cm hollow squares and the hand position was displayed as a 6-mm filled white square, the cursor spot. The position of the hand was used on-line to update the position of this cursor spot at 60 Hz. The relation between the actual hand location and the hand cursor spot was computer controlled so as to allow arbitrary visuomotor perturbations. Therefore the cursor spot could either accurately represent the true location of the hand or computer-controlled discrepancies between the cursor feedback and actual hand location could be introduced.

Paradigm. Subjects were asked to point to visually presented targets with their right hand. The experiment consisted of three parts: pre-exposure, exposure, and post-exposure. During pre- and post-exposure, subjects pointed to target T (10 repetitions for groups a–c; 15 repetitions for group d) in the absence of any visual feedback of the hand, starting from each of 7 starting locations (L1–L7; Fig. 2). This allowed the accuracy of pointing in the absence of visual feedback of hand location to be assessed for the 7 starting locations.

During the exposure phase, subjects repeatedly traced out a visual triangle L2–L6–T–L6–L2 forty times, thereby alternately pointing to the target from L2 and L6, while receiving feedback of hand location via the cursor spot. For the control group (Fig. 2a), the hand cursor spot accurately represented the actual hand position at all times. For the perturbation groups (Fig. 2b–d), displacements were surreptitiously introduced between the actual and visually displayed hand location. The displacement introduced increased linearly with distance from the starting location; the direction of the displacement varied between the groups. For movements made during the exposure phase the sign of the displacement was different for the two starting locations, L2 and L6. The dotted lines in Fig. 2b–d show the path taken by the visual feedback of the hand location and the solid lines the actual path taken by the hand. For example, for group b a discrepancy was introduced so that visual feedback of hand position was shifted to the left for movements made from L2, reaching a maximum

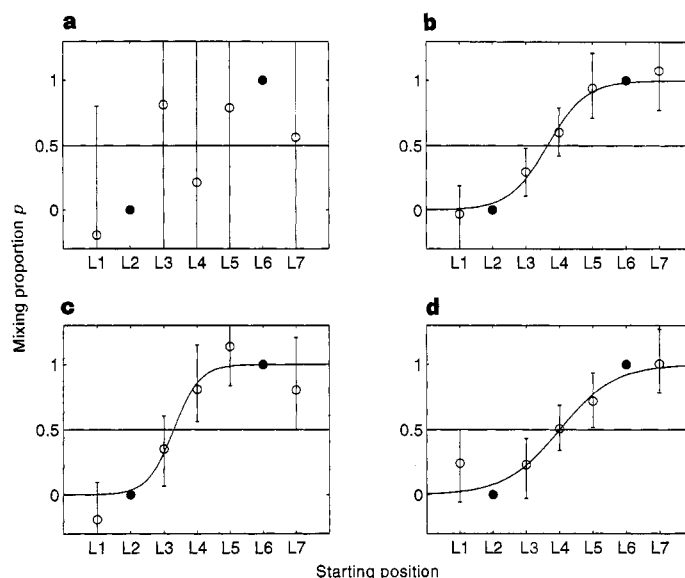


Figure 4 The mixing proportions with 95% confidence limits as a function of starting location for the **a**, control, and **b–d**, perturbation groups. For the i th starting location, the mixing proportion p is computed to minimize the distance between $\mathbf{v}_i$ and $p\mathbf{v}_1 + (1-p)\mathbf{v}_2$, where $\mathbf{v}_i$ is the mean adaptation vector for starting location L_i . Using this criterion the values of p are fixed to be 0 and 1 at starting locations L2 and L6 (indicated by filled circles), respectively. The values of p at points other than L2 and L6 capture the form of the generalization as a function of the two learned mappings at L2 and L6. Confidence intervals were computed on

this measure using bootstrap resampling³⁰. The mixing proportions for the control group (**a**) did not differ significantly from the null hypothesis of equal mixing (indicated by the line at 0.5). For groups **b–d**, a logistic function, $p(i) = 1/(1 + \exp(a + bi))$, representing the mixing probabilities predicted by the modular decomposition model (Fig. 1), was fitted to the mean mixing proportions (solid curve). All three fits were significant ($P < 0.001$) and the logistic function fit was a significantly better fit than linear regression over the ensemble data sets **b–d** ($F(15, 15) = 3.17$; $P = 0.02$).

discrepancy of 5 cm when the visual feedback of the hand was on target. For movements from L6, the visual feedback of hand position was shifted to the right, again reaching a maximum of 5 cm. Consequently, the single visual target location (T) was remapped to two distinct hand locations (P2 and P6 in Fig. 2), depending on whether the movement started from L2 or L6. Movements between L2 and L6 were unperturbed in all groups.

To assess learning and generalization to movements made from other starting locations, the subjects' change in pointing behaviour between the pre-exposure and post-exposure phases was analysed for each starting location. For each subject and start location, the average change in pointing position between the pre-exposure and post-exposure phases was calculated, together with the corresponding covariance matrices. The subjects' data were combined within each group for each starting location, obtaining the group mean change and the covariance matrix of the change for each starting location. The change in pointing from each starting location was plotted as a 95% confidence ellipse centred on the mean change (Fig. 3).

Received 23 August 1996; accepted 13 January 1997.

- Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. Adaptive mixture of local experts. *Neur. Comp.* 3, 79–87 (1991).
- Jordan, M. I. & Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neur. Comp.* 6, 181–214 (1994).
- Cacciatore, T. W. & Nowlan, S. J. Mixtures of controllers for jump linear and non-linear plants. In *Advances in Neural Information Processing Systems 6* (eds Cowan, J. D., Tesauero, G. & Alspector, J.) 719–726 (Morgan Kaufmann, San Francisco, 1994).
- Jacobs, R. A., Jordan, M. I. & Barto, A. G. Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks. *Cogn. Sci.* 15, 219–250 (1991).
- Graybiel, A. M., Aosaki, T., Flaherty, A. W. & Kimura, M. The basal ganglia and adaptive motor control. *Science* 265, 1826–1831 (1994).
- McGonigle, B. & Flook, J. Long-term retention of single and multistate prismatic adaptation by humans. *Nature* 272, 364–366 (1978).
- Welch, R. B., Bridgeman, B., Anand, S. & Browman, K. E. Alternating prism exposure causes dual adaptation and generalization to a novel displacement. *Percept. Psychophys.* 54, 195–204 (1993).
- Clower, D. M. et al. Role of posterior parietal cortex in the recalibration of visually guided reaching. *Nature* 383, 618–621 (1996).
- Kohler, I. Development and alterations of the perceptual world: conditioned sensations. *Proc. Austrian Acad. Sci.* 227, 1–118 (1951).
- Hay, J. C. & Pick, H. L. Gaze-contingent prism adaptation: optical and motor factors. *J. Exp. Psychol.* 72, 640–648 (1966).
- Shelhamer, M., Robinson, D. A. & Tan, H. S. Context-specific gain switching in the human vestibuloocular reflex. *Ann. NY Acad. Sci.* 656, 889–891 (1991).
- Baker, J. E., Perlmutter, S. I., Peterson, B. W., Rude, S. A. & Robinson, F. R. Simultaneous opposing adaptive changes in cat vestibulo-ocular reflex directions for two body orientations. *Exp. Brain Res.* 69, 220–224 (1987).
- Gandolfo, F., Mussa-Ivaldi, F. A. & Bizzi, E. Motor learning by field approximation. *Proc. Natl Acad. Sci. USA* 93, 3843–3846 (1996).
- Kravitz, J. H. & Yaffe, F. Conditioned adaptation to prismatic displacement with a tone as the conditional stimulus. *Percept. Psychophys.* 12, 305–308 (1972).
- Kravitz, J. H. Conditioned adaptation to prismatic displacement. *Percept. Psychophys.* 11, 38–42 (1972).
- Welch, R. B. Discriminative conditioning of prism adaptation. *Percept. Psychophys.* 10, 90–92 (1971).
- Martin, T. A., Keating, J. G., Goodkin, H. P., Bastian, A. J. & Thach, W. T. Throwing while looking through prisms. II. Specificity and storage of multiple gaze-throw calibrations. *Brain* 119, 1199–1211 (1996).
- Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. An internal model for sensorimotor integration. *Science* 269, 1880–1882 (1995).
- Andersen, R. A., Essek, C. & Siegel, R. Encoding of spatial location by posterior parietal neurons. *Science* 230, 456–458 (1985).
- Soechting, J. F. & Flanders, M. Sensorimotor representations for pointing to targets in three-dimensional space. *J. Neurophysiol.* 62, 582–594 (1989).
- Kalaska, J. F. & Crammond, D. J. Cerebral cortical mechanisms of reaching movements. *Science* 255, 1517–1523 (1992).
- Bridle, J. S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neuro-computing: Algorithms, Architectures, and Applications* (eds Fougelman-Soulie, F. & Hertz, J.) 227–236 (Springer, Berlin, 1990).
- Bedford, F. Constraints on learning new mappings between perceptual dimensions. *J. Exp. Psychol.* 15, 232–248 (1989).
- Imamizu, H., Uno, Y. & Kawato, M. Internal representations of the motor apparatus: implications from generalization in visuomotor learning. *J. Exp. Psychol.* 21, 1174–1198 (1995).
- Ghahramani, Z., Wolpert, D. M. & Jordan, M. I. Generalization to local remappings of the visuomotor coordinate transformation. *J. Neurosci.* 16, 7085–7096 (1996).
- Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. Neuronal population coding of movement direction. *Science* 233, 1416–1419 (1986).
- Georgopoulos, A. P. Current issues in directional motor control. *Trends Neurosci.* 18, 506–510 (1995).
- Bizzi, E., Mussa-Ivaldi, F. A. & Giszter, S. Computations underlying the execution of movement: A biological perspective. *Science* 253, 287–291 (1991).
- Wolpert, D. S., Ghahramani, Z. & Jordan, M. I. Are arm trajectories planned in kinematic or dynamic coordinates? An adaptation study. *Exp. Brain Res.* 103, 460–470 (1995).
- Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans* (Society for Industrial and Applied Mathematics, Philadelphia, 1982).

Acknowledgements. We thank M. I. Jordan for his support and G. E. Hinton, S. Goodbody and R. Lemon for comments on the manuscript. This project was supported by the Wellcome Trust, ATR Human Information Processing Research Laboratories, Siemens Corporation, the National Science Foundation, and the Office of Naval Research. Z.G. was supported by fellowships from the McDonnell-Pew Foundation and the Ontario Information Technology Research Centre.

Correspondence and requests for materials to Z.G. (e-mail: zoubin@cs.toronto.edu).

Protective effect of encapsulated cells producing neurotrophic factor CNTF in a monkey model of Huntington's disease

Dwaine F. Emerich^{*,†}, Shelley R. Winn[†], Philippe M. Hantraye[‡], Marc Peschanski[§], Er-Yun Chen[¶], Yaping Chu[¶], Patricia McDermott[¶], E. Edward Baetge^{*} & Jeffrey H. Kordower[¶]

^{*} CytoTherapeutics Inc., Providence, Rhode Island 02906, USA

[†] Department of Surgery, Oregon Health Sciences University, Portland, Oregon 97201, USA

[‡] URA CNRS 1285, Svc Hosp Frederic Joliet, 4 Place Gen Leclerc, Orsay 91406, France

[§] INSERM U421, Fac De Medicine, F94010 Creteil Cedex, France

[¶] Research Center for Brain Repair and Department of Neurological Sciences, Rush Presbyterian–St Luke's Medical Center, Chicago, Illinois 60612, USA

Huntington's disease is a genetic disorder that results from degeneration of striatal neurons, particularly those containing GABA (γ -aminobutyric acid)¹. There is no effective treatment for preventing or slowing this neuronal degeneration. Ciliary neurotrophic factor (CNTF) is a trophic factor for striatal neurons^{2,3} and therefore a potential therapeutic agent for Huntington's disease. Here we evaluate CNTF as a neuroprotective agent in a non-human primate model of Huntington's disease. We gave cynomolgus monkeys intrastriatal implants of polymer-encapsulated baby hamster kidney fibroblasts that had been genetically modified to secrete human CNTF. One week later, monkeys received unilateral injections of quinolinic acid into the previously implanted striatum to reproduce the neuropathology seen in Huntington's disease^{4,5}. Human CNTF was found to exert a neuroprotective effect on several populations of striatal cells, including GABAergic, cholinergic and diaphorase-positive neurons which were all destined to die following administration of quinolinic acid. Human CNTF also prevented the retrograde atrophy of layer V neurons in motor cortex and exerted a significant protective effect on the GABAergic innervation of the two important target fields of the striatal output neurons (the globus pallidus and pars reticulata of the substantia nigra). Our results show that human CNTF has a trophic influence on degenerating striatal neurons as well as on critical non-striatal regions such as the cerebral cortex, supporting the idea that human CNTF may help to prevent the degeneration of vulnerable striatal populations and cortical–striatal basal ganglia circuits in Huntington's disease.

Huntington's disease (HD) is an autosomal dominant neurodegenerative disease characterized by progressive movement disorder, with devastating psychiatric and cognitive deterioration⁶. Neuropathologically, HD is associated with a consistent and severe atrophy of the neostriatum. Although multiple populations of striatal neurons are affected in HD^{7,8}, the GABAergic medium-sized spiny projection neurons are particularly vulnerable¹. The cognitive and motor symptoms of HD result from the vulnerability of neostriatal neurons, possibly as a result of an endogenous disequilibrium of energy metabolism and excitotoxicity^{9–12}. There are no treatments available at present to slow neural degeneration in HD or for the behavioural symptoms that result from this cell loss. The administration of trophic factors may protect vulnerable

[¶] Present address: Alkermes, Inc., 64 Sidney Street, Cambridge, Massachusetts 02139, USA.

striatal neuronal populations in HD patients. CNTF is a member of the α -helical cytokine superfamily which functions in both the peripheral¹³⁻¹⁵ and central nervous system^{3,16-18}; it can prevent, in part, the loss of striatal neurons and protect against behavioural deficits in a rodent model of HD^{2,3}. Before clinical trials can be started, however, trophic factors need to be tested in a non-human primate model of HD.

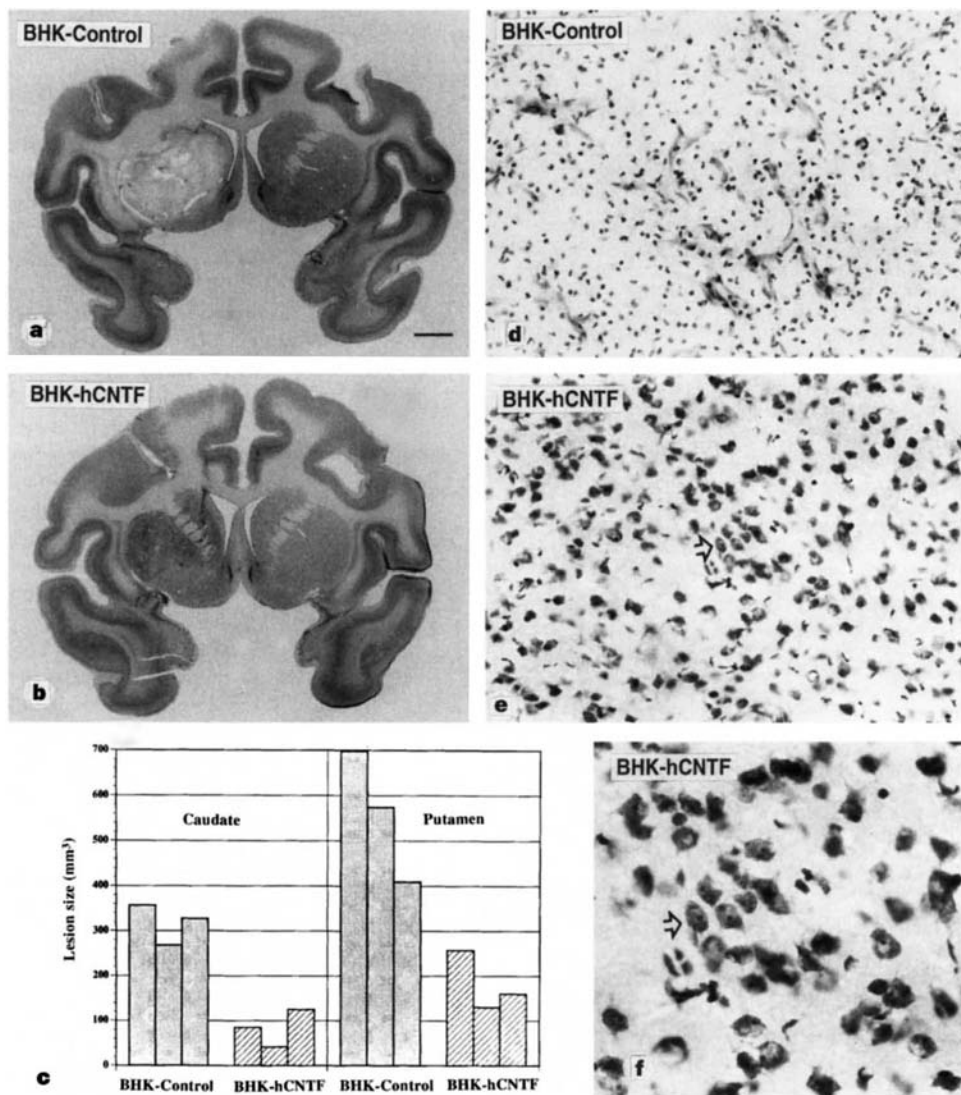
We used six cynomolgus monkeys in experiments that conformed to NIH guidelines. Three monkeys received unilateral intrastratial implants of polymer-encapsulated baby hamster fibroblasts which had been genetically modified to produce human CNTF (BHK-hCNTF)³. The remaining three monkeys served as controls and received identical implants except that the cells were not modified to produce hCNTF (BHK-control). One week later, all animals received injections of quinolinic acid (QA) into the ipsilateral caudate and putamen. All monkeys were killed 3 weeks later. Series of sections were stained for Nissl substance and the size of the lesion was quantified in all animals. Series of sections were also processed for immunocytochemical visualization of choline acetyltransferase (ChAT), glutamic acid decarboxylase (GAD), and dopamine and cyclic-AMP-regulated phosphoprotein (DARPP-32), as well as for histochemical staining of nicotinamide adenine dinucleotide phosphate diaphorase (NADPH-d)³. Neurons positive for diaphorase, ChAT and GAD in both the implanted/lesioned and intact striatum were counted as described³. The density of the GABAergic (DARPP-32) projections to the globus pallidus and

the pars reticulata of the substantia nigra were also quantified³. In addition, for each monkey, the size of neurons from layer V of motor cortex were quantified bilaterally from Nissl-stained sections³.

In the host striatum, QA induced a characteristic lesion of intrinsic neurons together with a substantial atrophy of the striatum. In monkeys implanted with BHK-control cells, Nissl-stained sections revealed extensive lesions in the caudate nucleus and putamen which were elliptical in shape. The lesion area in the three BHK-control animals averaged 317.72 mm³ (± 26.01) in the caudate and 560.56 mm³ (± 83.58) in the putamen. Many of the neurons that could be identified were shrunken and displayed a dystrophic morphology. In contrast, the size of the lesion was significantly reduced ($P < 0.05$) in BHK-hCNTF-implanted monkeys (Fig. 1). The lesion area in the three BHK-hCNTF monkeys averaged 83.72 mm³ (± 24.07) in the caudate and 182.54 mm³ (± 38.45) in the putamen. Many healthy Nissl-stained neurons were seen within the striatum of BHK-hCNTF-implanted monkeys after QA treatment even in regions near the needle tract (Fig. 1).

The neuroprotective effects of human CNTF were investigated by quantifying the loss of specific cell types within the striatum (Table 1). Lesioned monkeys receiving BHK-control implants had lost a significant number (caudate, 89.0%; putamen, 94.1%) of GAD-immunoreactive neurons within the striatum ipsilateral to the transplant. The excitotoxic degeneration of GAD-immunoreactive neurons in both the caudate nucleus and putamen was significantly attenuated by human CNTF as these monkeys had only a 64.0%

Figure 1 Low-power photomicrographs of Nissl-stained sections through the forebrain of monkeys receiving unilateral intracaudate and intraputamenal lesions with QA and encapsulated **a**, BHK-control, or **b**, BHK-hCNTF implants. Note the extensive lesion area on the left side which encompasses much of the striatum in **a**. In contrast, the damage in **b** is negligible for a monkey treated with hCNTF. **c**, The size of the lesion in both the caudate nucleus and putamen as quantified on Nissl-stained sections was significantly decreased in BHK-hCNTF-implanted monkeys (striped bars) relative to BHK-control-implanted animals (grey bars). Each bar represents data from an individual animal. **d**, **e**, Medium- and high-power (**f**) photomicrographs of Nissl-stained sections through the striatum of BHK-control- (**d**) and BHK-hCNTF-implanted (**e**, **f**) monkeys. **d**, A paucity of neurons is seen in the striatum of control-implanted monkeys. **e**, **f**, In contrast, numerous healthy neurons can be seen near a QA-injection site in BHK-hCNTF-implanted animals. For reference, the arrows in **e** and **f** point to the same location. Scale bar in **a** represents: in **a** and **b**, 4,000 μ m; in **d** and **e**, 50 μ m; in **f**, 25 μ m.



(caudate) and 64.1% (putamen) reduction in GAD-immunoreactive neurons in these regions relative to the contralateral side. As well as protecting the viability of GAD-immunoreactive striatal cells, human CNTF sustained the main striatal outflow pathways. The lesion-induced decrease in DARPP-32 immunoreactivity, a marker for GABAergic terminals, within the globus pallidus (Fig. 2a, b) and substantia nigra was prevented by BHK-hCNTF grafts. Optical density measurements revealed a significant reduction ($49.0 \pm 5.2\%$) in DARPP-32 immunoreactivity within the globus pallidus in BHK-control animals ipsilateral to the lesion. This reduction was significantly attenuated ($12.0 \pm 4.3\%$) in BHK-hCNTF implanted animals ($P < 0.05$). Likewise, the reduction in the optical density of DARPP-32 immunoreactivity in the pars reticulata of animals receiving BHK-control implants ($17 \pm 1.8\%$) was significantly attenuated ($4.3 \pm 0.4\%$) by human CNTF treatment ($P < 0.05$).

ChAT-immunoreactive and diaphorase-positive neurons were

also protected by cellular delivery of human CNTF (Table 1 and Fig. 2c, d). Monkeys receiving BHK-control implants had significantly reduced numbers of ChAT-positive neurons (Table 1 and Fig. 2e) within the caudate (80.3%) and putamen (87.9%), compared with only a 41.2% (caudate) and 54.1% (putamen) reduction in monkeys receiving BHK-hCNTF implants. Similarly, the loss of diaphorase-positive neurons in BHK-control-implanted monkeys (caudate, 83.9%; putamen, 86.6%) was significantly attenuated (caudate, 48.1%; putamen, 53.1%) in monkeys receiving encapsulated hCNTF implants (Fig. 2f, h).

QA produced a marked retrograde degeneration of cortical neurons in a region that projects to the striatum. Although neuron number was unaffected, monkeys receiving BHK-control implants had a significant atrophy (27%; $P < 0.05$) of neurons in layer V of the motor cortex ipsilateral to the lesion. This atrophy was significantly attenuated (6%; $P < 0.05$) in BHK-hCNTF-implanted

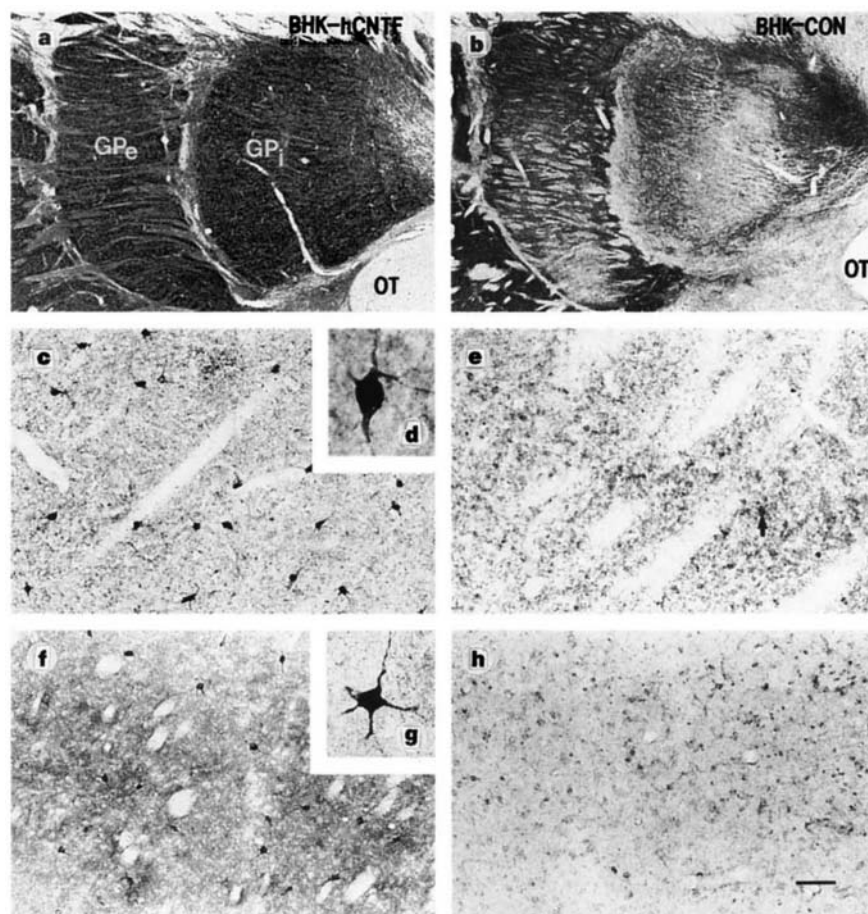
Table 1 Neuronal cell counts in QA-lesioned monkeys

	Caudate			Striatal region		
	Intact side	Lesioned/implanted side	Per cent loss	Intact side	Lesioned/implanted side	Per cent loss
GAD-positive neurons						
hCNTF	18,088 (437)	6,516 (1,233)	64.0*	17,731 (604)	6,374 (851)	64.1*
Control	19,298 (1,691)	2,133 (784)	89.0	16,674 (867)	978 (545)	94.1
ChAT-positive neurons						
hCNTF	2,259 (188)	1,328 (147)	41.2*	2,327 (209)	1,069 (36)	54.1*
Control	2,239 (164)	441 (189)	80.3	2,933 (172)	354 (253)	87.9
NADPH-d-positive neurons						
hCNTF	3,339 (553)	1,733 (489)	48.1*	3,508 (824)	1,647 (389)	53.1*
Control	3,405 (428)	548 (318)	83.9	3,393 (106)	458 (347)	86.5

Monkeys ($n = 3$ per group) receiving BHK-hCNTF implants displayed a significant rescue of striatal neurons in the QA-lesioned caudate and putamen. The number of neurons was compared across groups using a separate 2-way repeated measures ANOVA for each cell type.

* $P < 0.05$ for hCNTF compared with control.

Figure 2 a and b, Medium-power photomicrographs through the globus pallidus stained for DARPP-32 to visualize the GABAergic projection emanating from the striatum. **a**, In BHK-hCNTF-implanted monkeys, there is intense DARPP-32 immunoreactivity within both the external (GP_e) and internal (GP_i) segments of the globus pallidus, highlighting this projection which is sustained in these animals. **b**, In BHK-control-implanted animals, DARPP-32 immunoreactivity is reduced as a result of the QA lesion. The enhanced optical density of DARPP-32 immunoreactivity in the globus pallidus in BHK-hCNTF-implanted animals is statistically significant (see text for details) and reflects the significant implant-mediated sparing of GAD-immunoreactive neurons in these animals (Table 1). Many apparently healthy ChAT-immunoreactive neurons (**c, d**) and NADPH-d-positive (**f, g**) neurons are seen in the QA-lesioned striatum of monkeys receiving encapsulated BHK-hCNTF cells. In contrast, QA lesions resulted in a paucity of ChAT-immunoreactive (**e**) and NADPH-d-positive (**h**) neurons in BHK-control-implanted animals. Scale bar in **h** represents: in **a** and **b**, 500 μ m; in **c, e, f, h**, 100 μ m; in **d, g**, 25 μ m. OT, optic tract.



animals. Further analysis showed that the atrophy of cortical neurons was not due to general volumetric changes, but rather occurred preferentially in the medium-sized (300–400 μm and 400–500 μm cross-sectional area) neurons of the motor cortex (Fig. 3).

To evaluate the viability of implanted BHK cells and their expression of human CNTF *in vivo*, the implants were surgically retrieved just before death, 4 weeks after implantation. Implant devices were first assayed for human CNTF output by enzyme-linked immunosorbent assay (ELISA) and then embedded in glycol methacrylate for histological analysis³. Before implantation, BHK-hCNTF-implanted monkeys received capsules producing an average of 45.1 (± 4.1) ng human CNTF/24 h per capsule. At the time of their retrieval, the encapsulated cells secreted human CNTF at a rate of 12.7 (± 2.4) ng per 24 h per capsule. Histology indicated that viable BHK cells were evenly distributed at high density inside the capsule.

We have shown here that a therapeutic intervention can favourably influence the degeneration of striatal neurons and disruption of basal ganglia circuitry in a primate model of HD. Different strategies have demonstrated protection of striatal neurons from excitotoxicity and mitochondrial dysfunction^{19–23}, but none has provided direct evidence for protection of the GAD-immunoreactive neurons, the cell type most central in basal ganglia circuitry. Not only are GABAergic neurons viable in human-CNTF-treated monkeys, but DARPP-32 immunoreactivity reveals that the two critical GABAergic efferent projections from striatum to globus pallidus and the pars reticulata of the substantia nigra are sustained in these animals. Human CNTF implants also exert a strong neuroprotective effect on the cortical neurons innervating the striatum. Together, these data indicate that a major component of the basal ganglia loop circuitry, the cortical $\rightarrow$ striatal $\rightarrow$ globus pallidus/substantia nigra outflow circuitry is sustained by cellular delivery of human CNTF. Strategies directed towards preventing the degeneration of host systems have a special appeal for HD because potential sufferers can be positively identified well before the onset of symptoms and neural degeneration, even *in utero*. A successful approach should thus have a dramatic impact on the course of the disease. Our demonstration of neuroprotection by human CNTF should be extended by examining its behavioural effects in animal models of

HD and the results of its application both during and after neuronal degeneration. CNTF might be of great benefit because of its neuroprotective and symptomatic effects. Moreover, as excitotoxicity may influence such pathological conditions as ischaemia, and other neurodegenerative diseases such as Parkinson's and Alzheimer's^{24,25}, biologically delivered human CNTF may help prevent cell loss and its associated behavioural abnormalities in these disorders. $\square$

Methods

Cell preparation and encapsulation. The hCNTF gene was cloned into an expression vector containing a mutant dihydrofolate reductase gene (pNUT)³. The transcription of the hCNTF gene is driven by the metallothionein-1 promoter and uses the human growth hormone polyadenylation signal sequences for 3' processing. The 150-bp *XbaI*–*EcoRI* fragment of the immunoglobulin signal peptide was fused to the *EcoRI* site at the position of the natural hCNTF initiation codon to ensure secretion of the factor. The pNUT-hCNTF-TK construct was introduced into BHK cells using a standard calcium phosphate-mediated transfection method²⁶. BHK cells were grown and maintained as described³ and mock-transfected cells served as controls. BHK cells were encapsulated in asymmetric hollow fibres of poly(acrylonitrile-co-vinyl chloride) (PAN-PVC) copolymer as described^{3,27,28}. BHK cells plus or minus hCNTF were prepared as a single-cell suspension at a density of 2×10^7 cells per ml, mixed with collagen, and loaded into the capsules. All cell-loaded devices were assayed for hCNTF secretion by ELISA³.

Surgery. Monkeys were implanted with encapsulated BHK cells spaced 3–4 mm apart within the caudate ($n = 2$) and putamen ($n = 3$). Three received devices containing cells that were transfected to produce hCNTF, and 3 control monkeys received cells that had not been genetically modified. One week later, each monkey received 180 nmol QA in 10 μl injected midway between each polymer capsule, resulting in two injections in the putamen and one in the caudate.

Histology. Animals were transcardially perfused with phosphate-buffered saline (pH 7.4) followed by fixation with 3.5 l of 4% Zamboni's fixative. Frozen sections were cut (40 μm) on a sliding-knife microtome. Every sixth section through the striatum was processed immunocytochemically for ChAT (1:7,500), GAD (1:7,500) or diaphorase, using previously described procedures^{3,29}. The number of GAD-, ChAT- and diaphorase-positive neurons within the caudate and putamen were quantified throughout the striatum from 10 equispaced sections centred around the QA needle tract for both the caudate

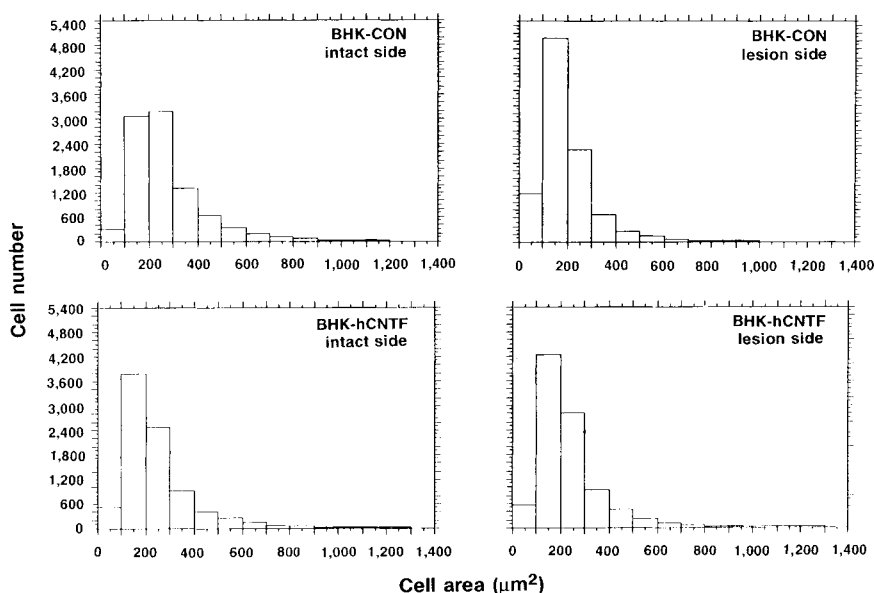


Figure 3 The distribution of layer V neurons in motor cortex based upon cell size, showing that hCNTF intrastratial grafts prevent atrophy. Note that in BHK-control animals, the increase in the number of neurons in the 0–100 μm and 100–200 μm range and the reduction of neurons in the 300–400 μm and 400–500 μm range

ipsilateral to the lesion. This shift in cell size was not seen in BHK-hCNTF-implanted monkeys. No significant changes were observed on the intact side between the BHK-control and BHK-hCNTF groups ($P > 0.05$; $n = 3$ animals per group).

and putamen. Every ChAT-, GAD-, or NADPH-d-positive neuron in each of these sections was counted bilaterally by an individual to whom the animal's experimental condition was unknown. The volume of the lesion produced by QA was determined in all animals using a semi-automated image analysis system (NIH Image 1200) as described³. Sections spaced 240 μ m apart that encompassed the entire lesion area were analysed. The border of the lesion was traced on Nissl-stained sections throughout each section and the volume of the lesioned area in each animal was expressed in cubic millimetres. The optical density of DARPP-32 within the globus pallidus was quantified bilaterally on seven sections per animal using the NIH image morphometric system. An additional series of 5 sections from the pars reticulata of the substantia nigra were analysed for optical density of DARPP-32-ir. The cross-sectional areas of all neurons within layer V were measured bilaterally in four Nissl-stained sections through the motor cortex in each animal. Over 37,000 neurons were quantified in these analyses.

Received 28 September 1996; accepted 27 January 1997.

- Reiner, A. *et al.* Differential loss of striatal projection neurons in Huntington's disease. *Proc. Natl Acad. Sci. USA* **85**, 5733–5737 (1988).
- Anderson, K. D., Panayotatos, N., Cordan, T. L., Lindsay, R. M. & Wiegand, S. J. Ciliary neurotrophic factor protects striatal output neurons in an animal model of Huntington's disease. *Proc. Natl Acad. Sci. USA* **93**, 7346–7351 (1996).
- Emerich, D. F. *et al.* Implants of encapsulated human CNTF-producing fibroblasts prevent behavioral deficits and striatal degeneration in a rodent model of Huntington's disease. *J. Neurosci.* **16**, 5168–5181 (1996).
- Beal, M. F. *et al.* Replication of the neurochemical characteristics Huntington's disease by quinolinic acid. *Nature* **321**, 168–171 (1986).
- Beal, M. F., Kowall, N. W., Swartz, K. J., Ferranti, R. J. & Martin, J. B. Differential sparing of somatostatin-neuropeptide Y and cholinergic neurons following striatal excitotoxin lesions. *Synapse* **3**, 38–47 (1989).
- Emerich, D. F. & Sanberg, P. R. in *Neuromethods: Animal Models of Neurological Disease* (eds Boulton, A. A., Baker, G. B. & Butterworth, R. F.) Vol. 17, 65–134 (Humana, New Jersey, 1992).
- Ferrante, R. J., Beal, M. F., Kowall, N. W., Richardson, E. P. & Martin, J. B. Sparing of acetylcholinesterase-containing striatal neurons in Huntington's disease. *Brain Res.* **415**, 178–182 (1987).
- Roberts, R. C. & DiFiglia, M. Short- and long-term survival of large neurons in the excitotoxic lesioned rat caudate nucleus: a light and electron microscopic study. *Synapse* **3**, 363–371 (1989).
- Albin, R. L. & Greenamyre, J. T. Alternative excitotoxic hypothesis. *Neurology* **42**, 733–738 (1992).
- Beal, M. F. Does impairment of energy metabolism result in excitotoxic neuronal death in neurodegenerative illnesses? *Ann. Neurol.* **31**, 119–130 (1992).
- Parker, W. D., Boyson, S. J., Luder, A. S. & Parks, J. K. Evidence for a defect on NADH:ubiquinone oxidoreductase (complex I) in Huntington's disease. *Neurology* **40**, 1231–1234 (1990).
- Wallace, D. C. W. Mitochondrial genetics: A paradigm for aging and degenerative diseases? *Science* **256**, 628–632 (1992).
- Apfel, S. C., Arezzo, J. C., Moran, M. & Kessler, J. A. Effects of administration of ciliary neurotrophic factor on normal motor and sensory peripheral nerves *in vivo*. *Brain Res.* **604**, 1–6 (1993).
- Sagot, Y. *et al.* Polymer encapsulated cell lines genetically engineered to release ciliary neurotrophic factor can slow down progressive motor neuronopathy in the mouse. *Eur. J. Neurosci.* **7**, 1313–1322 (1995).
- Sendtner, M. *et al.* Ciliary neurotrophic factor prevents degeneration of motor neurons in mouse mutant progressive motor neuronopathy. *Nature* **358**, 502–504 (1992).
- Clatterbuck, R. E., Price, D. L. & Koliatos, V. E. Ciliary neurotrophic factor prevents retrograde neuronal death in the adult central nervous system. *Proc. Natl Acad. Sci. USA* **90**, 2222–2226 (1993).
- Hagg, T., Quon, D., Higaki, J. & Varon, S. Ciliary neurotrophic factor prevents neuronal degeneration and promotes low affinity NGF receptor expression in the adult rat CNS. *Neuron* **8**, 145–158 (1993).
- Hagg, T. & Varon, S. Ciliary neurotrophic factor prevents degeneration of adult rat substantia nigra dopaminergic neurons *in vivo*. *Proc. Natl Acad. Sci. USA* **90**, 6315–6319 (1993).
- Emerich, D. F., Hammang, J. P., Baetge, E. E. & Winn, S. R. Implantation of polymer-encapsulated human nerve growth factor-secreting fibroblasts attenuates the behavioral and neuropathological consequences of quinolinic acid injections into rodent striatum. *Exper. Neurol.* **130**, 141–150 (1994).
- Frim, D. M. *et al.* Effects of biologically delivered NGF, BDNF and bFGF on striatal excitotoxic lesions. *Neuroreport* **4**, 67–370 (1993).
- Frim, D. M. *et al.* Striatal degeneration induced by mitochondrial blockade is prevented by biologically delivered NGF. *J. Neurosci. Res.* **35**, 452–458 (1993).
- Schumacher, J. M., Short, M. P., Hyman, B. T., Breakfield, X. O. & Isacson, O. Intracerebral implantation of nerve growth factor-producing fibroblasts protects striatum against neurotoxic levels of excitatory amino acids. *Neuroscience* **45**, 561–70 (1991).
- Tulipan, N., Luo, S.-Q., Allen, G. & Whetsell, W. O. Striatal grafts provide sustained protection from kainic and quinolinic acid-induced damage. *Exp. Neurol.* **102**, 325–332 (1988).
- Choi, D. W. Glutamate neurotoxicity and diseases of the nervous system. *Neuron* **1**, 623–634 (1988).
- Olney, J. W. Excitatory amino acids and neuropsychiatric disorders. *Biol. Psychiatry* **26**, 505–525 (1989).
- Baetge, E. E., Suh, Y. H. & Joh, T. H. Complete nucleotide and deduced amino acid sequence of bovine phenylethanolamine N-methyltransferase: partial homology with rat tyrosine hydroxylase. *Proc. Natl Acad. Sci. USA* **83**, 5454–5458 (1986).
- Winn, S. R. *et al.* Polymer-encapsulated cells genetically modified to secrete human nerve growth factor promote the survival of axotomized septal cholinergic neurons. *Proc. Natl Acad. Sci. USA* **91**, 23–28 (1994).
- Emerich, D. F. *et al.* Transplantation of polymer-encapsulated cells genetically modified to secrete human nerve growth factor prevents the loss of degenerating cholinergic neurons in nonhuman primates. *J. Comp. Neurol.* **349**, 148–164 (1994).
- Kordower, J. H., Chen, E.-Y., Mufson, E. J., Winn, S. R. & Emerich, D. F. Intrastriatal implants of polymer-encapsulated cells genetically modified to secrete human NGF: trophic effects upon cholinergic and noncholinergic neurons. *Neuroscience* **72**, 63–77 (1996).

Acknowledgements. We thank P. Greengard for the DARPP-32 antibody and L. Martel for histological assistance. This work was supported by an NIH grant to J.H.K.

Correspondence and requests for materials should be addressed to D.F.E.

The *Pax4* gene is essential for differentiation of insulin-producing β cells in the mammalian pancreas

Beatriz Sosa-Pineda*, Kamal Chowdhury, Miguel Torres*, Guillermo Oliver* & Peter Gruss

Max Planck Institute for Biophysical Chemistry, Am Fassberg 37077 Göttingen, Germany

The mammalian pancreas contains two distinct cell populations: endocrine cells which secrete hormones into the bloodstream, and exocrine cells, which secrete enzymes into the digestive tract¹. The four endocrine cell types found in the adult pancreas— α , β , δ and PP—synthesize glucagon, insulin, somatostatin and pancreatic polypeptide, respectively². All of these endocrine cells arise from common multipotent precursors, which coexpress several hormones when they start to differentiate³. Expression of some homeobox genes in the early developing pancreas has been reported^{4–7}. The *Pax4* gene is expressed in the early pancreas, but is later restricted to β cells. Inactivation of *Pax4* by homologous recombination results in the absence of mature insulin- and somatostatin-producing cells (β and δ , respectively) in the pancreas of *Pax4* homozygous mutant mice, but glucagon-producing α cells are present in considerably higher numbers. We propose that the early expression of *Pax4* in a subset of endocrine progenitors is essential for the differentiation of the β and δ cell lineages. A default pathway would explain the elevated number of α cells in the absence of *Pax4*.

A genomic screen has previously identified *Pax4* as a member of the murine family of *Pax* genes⁸. By using reverse transcription-polymerase chain reaction (RT-PCR) we have isolated the corresponding cDNA sequence. *Pax4* mRNA was detected in a few cells in the ventral spinal cord and pancreas as early as day 9.5 of development (E9.5) (B.S.-P., M. Asano & P.G., manuscript in preparation). To investigate the function of *Pax4* during development, we inactivated this gene by homologous recombination in embryonic stem (ES) cells. Almost the entire paired box domain was deleted and replaced by the β -galactosidase gene fused in frame with the amino terminus of *Pax4* (Fig. 1a). Detection of LacZ activity allowed a detailed analysis of *Pax4* expression throughout development.

Heterozygous *Pax4* offspring did not exhibit any obvious abnormalities, survived to adulthood, and were fertile. *Pax4*-deficient offsprings were born with the expected Mendelian distribution, indicating that the absence of *Pax4* is not lethal *in utero*. Newborn *Pax4*^{−/−} mice appeared normal and were indistinguishable from their littermates. However, 48 h later they exhibited growth retardation and dehydration (Fig. 1c), and died within three days of birth, demonstrating the complete penetrance of the mutant phenotype. These abnormalities resembled those for *Pdx1*^{−/−} mice which lack a pancreas^{9,10}. However, in *Pax4*-deficient mice the pancreas is present and exhibits a normal macroscopic appearance.

A comparative analysis of the expression of LacZ activity was then performed in pancreas from *Pax4*^{+/−} and *Pax4*^{−/−} embryos. Staining of heterozygous embryos showed that there was LacZ activity in cells within the dorsal pancreas at around E10.5 (Fig. 2a), and in the ventral pancreas around E11.0 (data not shown). In *Pax4*^{+/−} newborn pancreas, LacZ activity was restricted to discrete areas that coexpressed insulin, indicating that they

* Present addresses: Department of Genetics, St Jude Children's Research Hospital, 332 North Lauderdale, P.O. Box 318, Memphis, Tennessee 38101-0318, USA (B.S.-P. and G.O.); Departamento de Inmunología y Oncología, Centro Nacional de Biotecnología, Universidad Autónoma, Madrid 28049, Spain (M.T.).

correspond to the insulin-producing β cells of the islets of Langerhans (Fig. 2b). Glucagon-producing α cells were found with their typical organization at the periphery of the insulin core, and did not show LacZ activity (Fig. 2c). From E10.5 no differences in LacZ expression were observed in the pancreas of $Pax4^{+/+}$ and $Pax4^{-/-}$ embryos (compare Fig. 2a and d) until approximately E16.5, when

this activity was diminished in homozygous pancreas (data not shown). LacZ expression was undetectable at birth in $Pax4^{-/-}$ pancreas (compare Fig. 2c and e).

To study possible alterations in the pancreatic cell population, we analysed the expression of insulin and glucagon. Early in development (E10.5), insulin- and glucagon-producing cells were detected

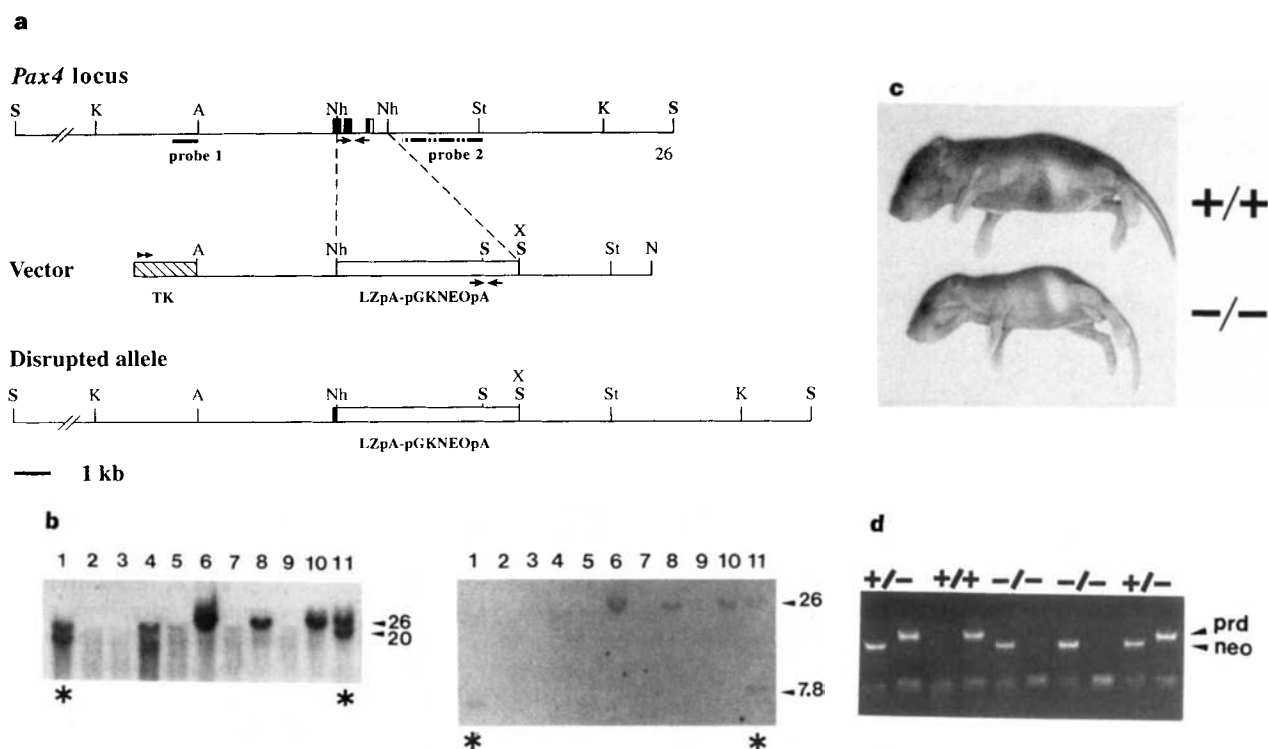


Figure 1 Targeted disruption of *Pax4* and generation of *Pax4*^{-/-} mice. **a**, Structure of wild-type and targeted *Pax4* loci. Targeted deletion of almost all the *Pax4* paired box domain (dark boxes; exons 2, 3 and 4) was produced, by fusing in frame a β -galactosidase-neomycin resistance cassette (double arrowheads indicate direction of transcription). Restriction enzymes: A, *Apa*I; K, *Kpn*I; Nh, *Nhe*I; S, *Spe*I; St, *Stu*I; X, *Xho*I. **b**, DNA isolated from ES cell clones was digested with *Spe*I and analyzed by Southern blot hybridization with a 0.7-kb external genomic fragment (probe 1, left), and a 0.5-kb cDNA fragment (probe 2, right). Sizes are in kilobases.

In both, a 26-kb fragment is indicative of the wild-type allele; the 20-kb and 7.8-kb, respectively, originate from the targeted allele. Asterisks indicate two clones positive for homologous recombination. **c**, Littermates 2 days old, wild-type (top) and *Pax4*-deficient (bottom), showing the reduced size of null-mutant mice. **d**, Embryos were genotyped by PCR analysis of genomic DNA from yolk sacs or tails, and two sets of primers (indicated by arrows in **a**) were used to amplify wild-type paired (prd) domain and/or the neo gene.

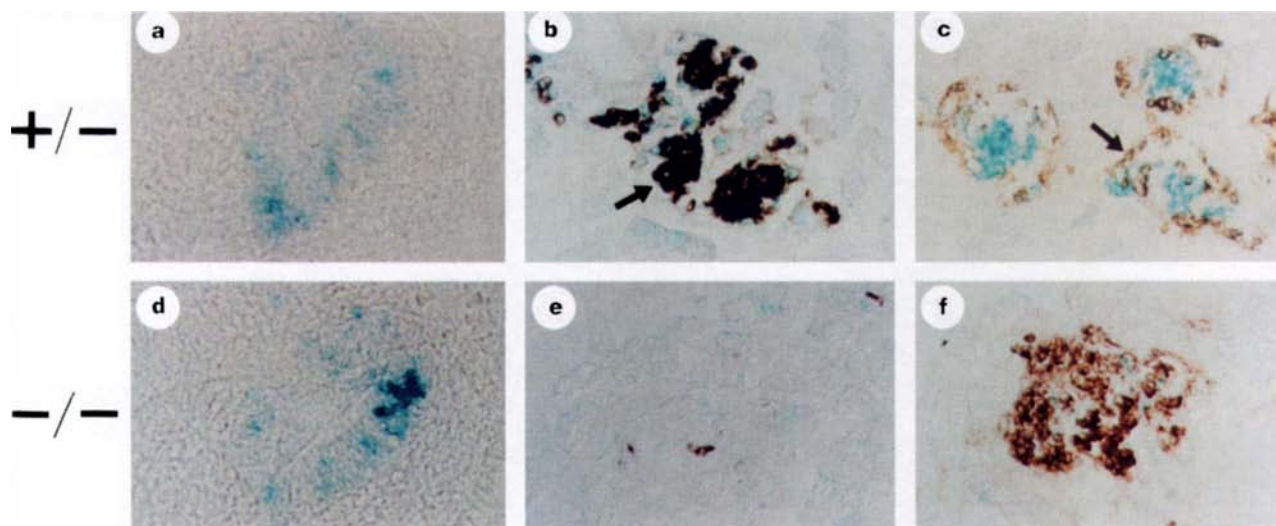


Figure 2 Analysis of LacZ, insulin and glucagon expression in the pancreas of *Pax4*^{+/+} and *Pax4*^{-/-} mice. Initial expression (E10.5) of LacZ activity within dorsal pancreas seem indistinguishable between *Pax4*^{+/+} and *Pax4*^{-/-} embryos (compare **a** and **d**). At birth, LacZ expression in *Pax4*^{+/+} pancreas is restricted to a discrete region that coexpresses insulin (arrow in **b**), corresponding to β cells. In *Pax4*^{-/-} newborn pancreas, however, LacZ activity is barely detected, and few if any insulin

cells are present (compare **b** and **e**). In *Pax4*^{+/+} newborn pancreas, the less numerous glucagon (α) cells surround the β -cell core and show no LacZ activity (arrow in **c**). Instead, *Pax4*^{-/-} newborn pancreas contains a larger number of glucagon cells, which are abnormally clustered (see **f**). **a** and **d**, Vibratome sections from LacZ-stained E10.5 embryos. **b**, **c**, **e** and **f**, Paraffin sections from LacZ-stained newborn pancreas, subsequently assayed for immunohistochemistry. Magnification, $\times 400$.

in the pancreas of both *Pax4*^{+/+} and *Pax4*^{-/-} embryos (compare Fig. 3 and d, and 3b and e). However, few if any insulin-producing cells were found in the pancreas of *Pax4*^{-/-} newborn mice (Fig. 2e), whereas glucagon-cells were numerous and abnormally clustered (compare Fig. 2c and f). In normal newborn pancreas, most of the endocrine cells are β cells, with α cells representing only a small fraction¹. These data suggest that, in the absence of functional *Pax4*, maturation of the pancreatic β cells was affected. To confirm this observation we analysed the expression of Pdx1, a marker specific for pancreatic progenitors and mature β cells. During murine pancreas development the homeobox containing gene *Pdx 1* starts to be expressed in the earliest progenitors, but later becomes restricted to fully differentiated β cells¹¹⁻¹³. At E10.5 we found no differences in Pdx1 expression in the pancreas of *Pax4*^{+/+} and *Pax4*^{-/-} embryos (compare Fig. 3c and f). However, in *Pax4*^{-/-} newborn pancreas, Pdx1 expression was not detected (compare Fig. 4a and b). This confirms our initial observation that mature β cells are absent in *Pax4*^{-/-} mice.

We then investigated the expression of somatostatin, another endocrine marker specific for δ cells. These cells start to differentiate around E14.5 in murine pancreas, rather later than α and β cells¹⁴. In normal newborn pancreas, δ cells are found in small numbers, mainly distributed in the periphery of the islets of Langerhans and intermingled with α cells (Fig. 4c). In *Pax4*^{-/-} newborn pancreas no somatostatin expression was detected (Fig. 4d), suggesting that inactivation of *Pax4* also affects differentiation of the δ lineage. These results are not really surprising as it is probable that β and δ cells arise from a common precursor^{3,14} that diverges around E14.5 in mouse development. Expression of somatostatin in the gut seemed to be unaffected (data not shown).

In the exocrine pancreas a variety of digestive enzymes are stored within secretory granules. Most of these are secreted as inactive precursors, which are in turn activated in the duodenum². When the newborns start suckling milk, digestive enzymes become depleted in the exocrine cells of pancreas¹⁵. To analyse possible modifications in exocrine pancreatic tissue from *Pax4*^{-/-} mice, expression of the enzyme α -amylase was monitored. When pancreas from 3 days old wild-type mice was analysed, large portions of tissue showed little or no expression of α -amylase (Fig. 4e). We know that *Pax4*-deficient mice are able to suckle at birth as their stomachs are full with milk

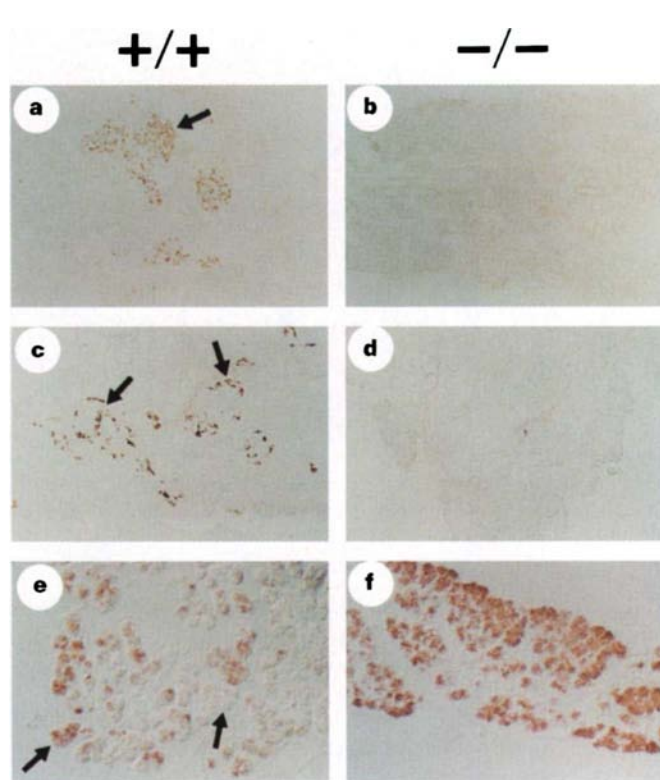


Figure 4 Expression of Pdx1, somatostatin and α -amylase in *Pax4*^{+/+} and *Pax4*^{-/-} pancreas from newborn mice. In the endocrine *Pax4*^{+/+} newborn pancreas, Pdx1 expression is restricted to insulin-producing (β) cells (arrow in **a**), whereas somatostatin (δ) cells are mainly located in the periphery (arrows in **c**) of the islets of Langerhans. Neither protein is detected in pancreas from newborn *Pax4*^{-/-} mice (compare **a** with **b**, and **c** with **d**). In exocrine *Pax4*^{+/+} newborn pancreas the content of α -amylase within the cells varies (arrows in **e**), as a result of the depletion of digestive enzymes that normally follows suckling. In almost all the exocrine tissue of newborn *Pax4*^{-/-} pancreas, however, a large amount of α -amylase is present (compare **e** and **f**), suggesting that its secretion may be altered. Magnification, $\times 200$.

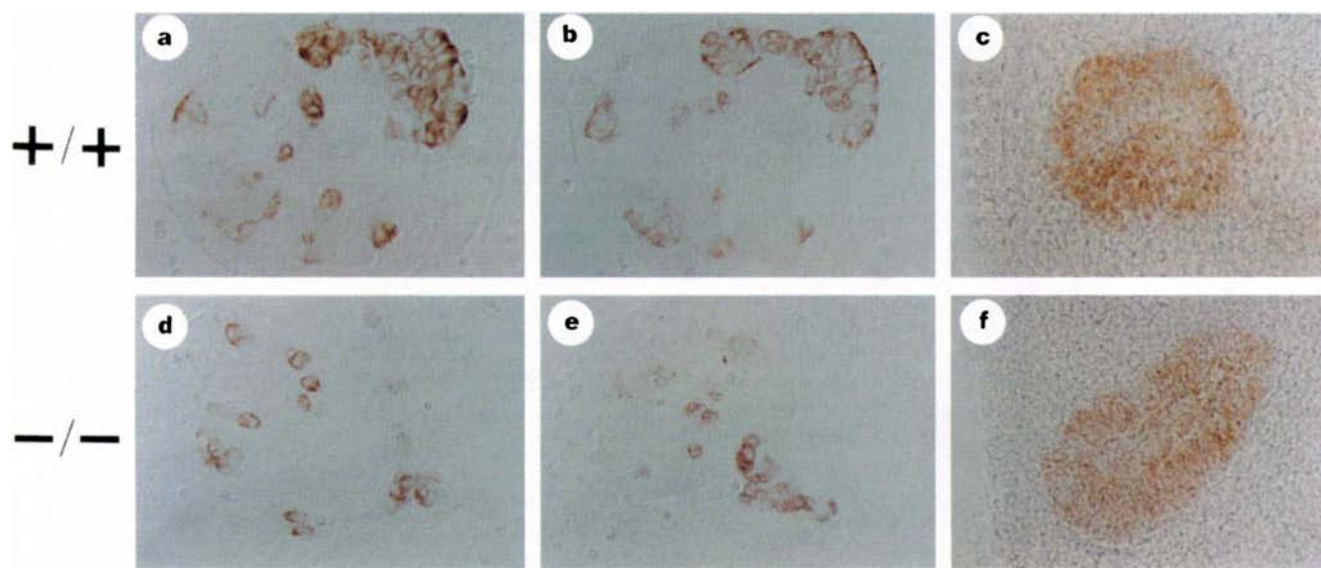


Figure 3 Expression of insulin, glucagon and Pdx1 in the early (E10.5) developing pancreas from *Pax4*^{+/+} and *Pax4*^{-/-} embryos. **a, b**, Adjacent sections from dorsal *Pax4*^{+/+} pancreas showing expression of insulin and glucagon, respectively. Coexpression of both hormones by undifferentiated precursors has been reported at this stage^{14,16}. Insulin- and glucagon-producing cells are also detected

in the pancreas of *Pax4*^{-/-} E10.5 embryos (**d** and **e**, respectively). Expression of Pdx1 in the dorsal pancreas from *Pax4*^{+/+} and *Pax4*^{-/-} E10.5 embryos also seems indistinguishable (compare **c** and **f**). **a, b, d** and **e**, Paraffin sections. **c** and **f**, Vibratome sections. Magnification, $\times 400$.

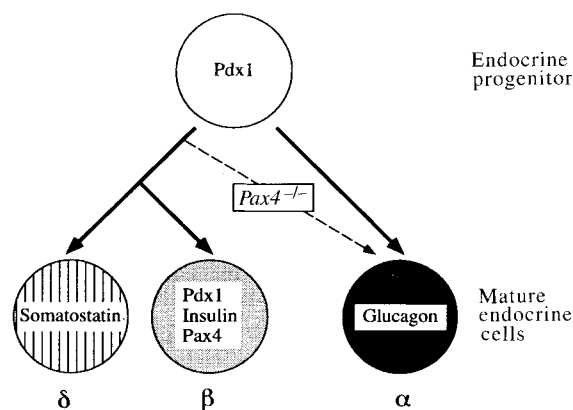


Figure 5 Model for the role of *Pax4* in endocrine differentiation in the mouse pancreas. Earliest endocrine progenitor cells are characterized by their expression of *Pdx1* (ref. 11) in mature endocrine cells, *Pdx1*, insulin and *Pax4* are specifically expressed in the β lineage.

(Fig. 1c). Surprisingly, *Pax4*^{-/-} pancreas at 3 days old showed a strong expression of α -amylase in all exocrine cells, suggesting that they might not be able to secrete their enzymes into the digestive tract (Fig. 4f).

Endocrine cell differentiation in the developing pancreas of the mouse has been extensively studied^{3,11,14}. Identification of precursor cells and their descendants has been achieved based on their coexpression of different markers. Accordingly (see Fig. 5), a common endocrine progenitor expressing *Pdx1* appears around E8.5 in a restricted area of mouse endoderm¹¹. Earliest insulin (β) and glucagon (α) precursors are present at E9.5, and they show coexpression of both hormones^{14,16} within the next couple of days. Differentiation of the α lineage occurs earlier in development in precursors that contain glucagon but presumably had lost expression of *Pdx1*. The earliest β -cell progenitors contain *Pdx1* and insulin. Maintenance of their expression is thought to be important if they are to attain full β -cell maturation¹⁷.

In the pancreas of *Pax4*^{-/-} embryos, insulin, glucagon and *Pdx1* are present, indicating that *Pax4* is not required for generation of the earliest α and β precursors. However, the onset of expression of *Pax4* during pancreas development (around E9.5) and the absence of mature β and δ cells in pancreas of *Pax4*^{-/-} newborn mice suggest an early role in the initial processes of endocrine differentiation (see Fig. 5). This is supported by the observation that insulin is no longer detected in *Pax4*^{-/-} pancreas at E13.5, a stage when β cells are not expected to be fully differentiated (data not shown). It is important to identify those cells initially expressing *Pax4* to gain an insight into the early role of *Pax4* in pancreas development, and anti-*Pax4* antibodies seem useful. The appearance of a larger number of glucagon cells in the absence of *Pax4* suggests that, in *Pax4*^{-/-} pancreas, a significant proportion of early endocrine precursors had lost their initial commitment to become β cells, acquiring instead an α -cell phenotype (Fig. 5, broken arrow). It remains possible, however, that the absence of β and δ cells results in an increased proliferation of α -cell precursors. In the differentiated endocrine pancreas, *Pax4* expression seems to be restricted largely to β cells, where they persist until adulthood (data not shown). This suggests that later in development, *Pax4* activity may contribute to maintain the β -cell population in a differentiated state. □

Methods

Preparation of construct for homologous recombination and generation of *Pax4*-deficient mice. Targeted deletion of almost all of the paired box domain (exons 2–4) of *Pax4* was achieved by fusing in frame a β -galactosidase-neomycin resistance cassette. A 5.1-kb *Xba*–*Xho*I fragment containing LacZpA-pGKNeopA sequences was blunt-ended and ligated into the *Nhe*I-

digested and blunt-ended *Pax4* construct. A 5' *Nhe*I site was retained and a new *Spe*I site was generated at the 3' end. An F9 polyoma early promoter-derived HSV TK gene was added upstream of the 5' homology for negative selection. R1 ES cells were electroporated and selected following standard procedures. Positive clones were used to generate chimaeras by morula aggregation. Tail and yolk-sac DNA isolation and PCR amplification of genomic DNA were performed. The phenotype was analysed in NMR1 and C57BL/6 mice; no differences were observed.

Detection of LacZ activity in embryos and tissues. For detection of LacZ activity, X-gal staining of mouse embryos and isolated newborn pancreas were performed. A post-fixation in 4% paraformaldehyde at 4 °C was done overnight. For immunohistochemistry, X-gal-stained embryos and tissues were embedded in paraffin and sectioned (10 μ M) with a microtome.

Immunohistochemistry on sections. Immunohistochemistry was performed on paraffin sections¹⁸ and cryostat sections¹², with the following modifications: newborn pancreas was fixed 3 h with 4% paraformaldehyde, and cryoprotected with 30% sucrose in PBS overnight. After washing off the secondary antibody, sections were incubated with 0.6% H₂O₂ for 20 min.

Whole-mount immunohistochemistry. Whole-mount immunostaining of E10.5 mouse embryos was performed as described¹⁹ except that bleaching was for 24 h. After staining, E10.5 embryos were post-fixed, embedded in a gelatin-BSA matrix, and sectioned 30 μ m with a vibratome.

Antibodies. Primary antibodies were: monoclonal anti-insulin and anti-glucagon (Sigma); rabbit anti-PDX1 (IPF1; provided by H. Edlund); rabbit anti-PDX1 (provided by C. Wright); rabbit anti-somatostatin (Genosys); mouse anti α -amylase antibody (Sigma). Anti-mouse antibody coupled to HRP (Capell) was detected with a colour-development substrate solution containing diaminobenzidine. Anti-rabbit antibody coupled to HRP (Capell) was used to detect anti-PDX1 antibodies. Anti-rabbit antibody coupled to biotin was used to detect anti-somatostatin and anti-PDX1 antibodies in newborn pancreas, and was detected with the ABC immunoperoxidase system (Vector).

Received 14 October 1996; accepted 24 January 1997.

- Slack, J. M. W. Developmental biology of the pancreas. *Development* **121**, 1569–1580 (1995).
- Pictet, R. L. An ultrastructural analysis of the developing embryonic pancreas. *Dev. Biol.* **29**, 436–467 (1972).
- Alpert, S., Hanahan, D. & Teitelman, G. Hybrid insulin genes reveal a developmental lineage for pancreatic endocrine cells and imply a relationship with neurons. *Cell* **53**, 295–308 (1988).
- Oliver, G. et al. *Pax4*, a prospective homeobox gene expressed during mouse development. *Mech. Dev.* **44**, 3–16 (1993).
- Raju, K. et al. Characterization and developmental expression of *Tlx-1*, the murine homolog of *Hox11*. *Mech. Dev.* **44**, 51–64 (1993).
- Turque, N., Plaza, S., Radvanyi, F., Carriere, C. & Saule, S. *Pax-QNR/Pax-6*, a paired box- and homeobox-containing gene expressed in neurons, is also expressed in pancreatic endocrine cells. *Mol. Endocrinol.* **8**, 929–938 (1994).
- Wright, C. V. E., Schneegelsberg, P. & De Robertis, E. M. *XIH8* box 8: a novel *Xenopus* homeo protein restricted to a narrow band of endoderm. *Development* **104**, 787–794 (1988).
- Walther, C. et al. *Pax*: a murine multigene family of paired box-containing genes. *Genomics* **11**, 424–434 (1991).
- Jonsson, J., Carlsson, L., Edlund, T. & Edlund, H. Insulin-promoter factor 1 is required for pancreas development in mice. *Nature* **371**, 606–609 (1994).
- Offield, M. F. et al. *PDX-1* is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* **122**, 983–995 (1996).
- Guz, Y. et al. Expression of murine *STF-1*, a putative insulin gene transcription factor, in β -cells of pancreas, duodenal epithelium and pancreatic exocrine and endocrine progenitors during ontogeny. *Development* **121**, 11–18 (1995).
- Ohlsson, H., Carlsson, K. & Edlund, T. IPF1, a homeo-domain containing transactivator of the insulin gene. *EMBO J.* **13**, 1145–1158 (1994).
- Miller, C. P., McGehee, R. E. & Habener, J. F. *IDX-1*: a new homeodomain transcription factor expressed in rat pancreatic islets and duodenum that transactivates the somatostatin gene. *EMBO J.* **13**, 1145–1156 (1994).
- Teitelman, G., Alpert, S., Polak, J. M., Martinez, A. & Hanahan, D. Precursor cells of mouse endocrine pancreas coexpress insulin, glucagon and the neuronal proteins tyrosine hydroxylase and neuropeptide Y, but not pancreatic polypeptide. *Development* **118**, 1031–1039 (1993).
- Githens, S. in *The Pancreas: Biology, Pathobiology, and Disease* 2nd edn (ed. V. L. W. Go et al.) 21–73 (Raven, New York, 1993).
- Gittes, G. K. & Rutter, W. J. Onset of cell-specific gene expression in the developing mouse pancreas. *Proc. Natl. Acad. Sci. USA* **89**, 1128–1132 (1992).
- Serup, P. et al. Induction of insulin and islet amyloid polypeptide production in pancreatic islet glucagonoma cells by insulin promoter factor 1. *Proc. Natl. Acad. Sci. USA* **93**, 9015–9020 (1996).
- Oliver, G., Wright, C. V. E., Hardwicke, J. & De Robertis, E. M. Differential antero-posterior expression of two proteins encoded by a homeobox gene in *Xenopus* and mouse embryos. *EMBO J.* **7**, 3199–3209 (1988). (Author: Vol. no. OK?)
- Dent, J. A., Polson, A. G. & Klymkowsky, M. W. A whole-mount immunocytochemical analysis of the expression of the intermediate filament protein vimentin in *Xenopus*. *Development* **105**, 61–74 (1989).

Acknowledgements. We thank H. Edlund and C. Wright for the anti-Pdx1 (IPF1) antibody; H. Edlund for advice with immunohistochemistry in cryostat sections; M. Kessel, L. St-Onge and A. Stoykova for critical reading of the manuscript; P. Geisendorff, S. Mahsur, M. Walther and R. Scholz for technical assistance; and R. Altschaffel for photographic artwork. This work was supported by the Max-Planck Society.

Correspondence and requests for material should be addressed to P.G. (e-mail: pgrass@gwdg.de). The sequence for *Pax4* has been deposited in the EMBL nucleotide sequence database, accession no. Y09584.

Abnormal angiogenesis and responses to glucose and oxygen deprivation in mice lacking the protein ARNT

Emin Maltepe*, Jennifer V. Schmidt†‡, David Baunoch*, Christopher A. Bradfield§ & M. Celeste Simon||§#

Departments of * Pathology, || Medicine, ‡ Molecular Genetics and Cell Biology, # Howard Hughes Medical Institute, 5841 S. Maryland Avenue, MC 1028, University of Chicago, Chicago, Illinois 60637, USA

† Department of Molecular Pharmacology and Biological Chemistry, Northwestern University Medical School, Chicago, Illinois 60611, USA

§ McArdle Laboratory for Cancer Research, University of Wisconsin, Madison, Wisconsin 53706, USA

The arylhydrocarbon-receptor nuclear translocator (ARNT) is a member of the basic-helix-loop-helix-PAS family of heterodimeric transcription factors which includes the arylhydrocarbon receptor (AHR), hypoxia-inducible factor-1 α (HIF-1 α) and the *Drosophila* single-minded protein (Sim)¹⁻⁴. ARNT forms heterodimeric complexes with the arylhydrocarbon receptor, HIF-1 α , Sim and the PAS protein Per^{2,4-6}. In response to environmental pollutants, AHR-ARNT heterodimers regulate genes involved in the metabolism of xenobiotics⁷⁻⁹, whereas ARNT-HIF-1 α heterodimers probably regulate those involved in the response to oxygen deprivation¹⁰⁻¹³. By generating a targeted disruption of the *Arnt* locus in the mouse, we show here that *Arnt*^{-/-} embryonic stem cells fail to activate genes that normally respond to low oxygen tension. *Arnt*^{-/-} ES cells also failed to respond to a decrease in

glucose concentration, indicating that ARNT is crucial in the response to hypoxia and to hypoglycaemia. *Arnt*^{-/-} embryos were not viable past embryonic day 10.5 and showed defective angiogenesis of the yolk sac and branchial arches, stunted development and embryo wasting. The defect in blood vessel formation in *Arnt*^{-/-} yolk sacs is similar to the angiogenic abnormalities reported for mice deficient in vascular endothelial growth factor^{14,15} or tissue factor¹⁶. On the basis of these findings, we propose a model in which increasing tissue mass during organogenesis leads to the formation of hypoxic/nutrient-deprived cells, the subsequent activation of ARNT, and a concomitant increase in the expression of genes (including that encoding vascular endothelial growth factor) that promote vascularization of the developing yolk sac and solid tissues.

The exon encoding the basic-helix-loop-helix (bHLH) domain of the murine *Arnt* gene was disrupted by insertion of the phosphoglycerate kinase (PGK) promoter/neomycin-resistance complementary DNA (Fig. 1a). Heterozygous embryonic stem (ES) cells were generated by homologous recombination (Fig. 1b, lanes 2 and 3) and then used to obtain *Arnt*^{-/-} ES cells through high G418 selection¹⁷ (Fig. 1b, lanes 4 and 5). *Arnt*^{+/-} cells expressed ARNT protein at about half the level of *Arnt*^{+/+} cells (Fig. 1c, lanes 2 and 3). *Arnt*^{-/-} ES cells did not produce detectable amounts of ARNT messenger RNA or protein (Fig. 1c, lanes 4 and 5, and Fig. 2a, lanes 5-12).

ARNT and HIF-1 α are expressed in R1 ES cells (Fig. 2a). We therefore tested the requirement for ARNT in the regulation of hypoxia-responsive genes under normal and hypoxic conditions (1.5% O₂, normoxia being 21%). In response to hypoxia, many tissues switch from aerobic to anaerobic metabolism through activation of glycolytic pathways, increase their glucose uptake to compensate for the reduced amount of ATP generated in glycolysis, and increase local vascularization by stimulating angiogenesis. The glycolytic enzymes phosphoglycerate kinase (PGK-1) and aldolase A (ALDA), the glucose transporter GLUT-1 and the endothelial-cell-specific mitogen vascular endothelial growth factor (VEGF), are all critical for the hypoxic response and are encoded by genes that are activated by hypoxia¹⁸⁻²⁰. *Arnt*^{+/+} ES cells cultured in 1.5% O₂

* Present address: Howard Hughes Medical Institute, Princeton University, Princeton, New Jersey, 08544, USA.

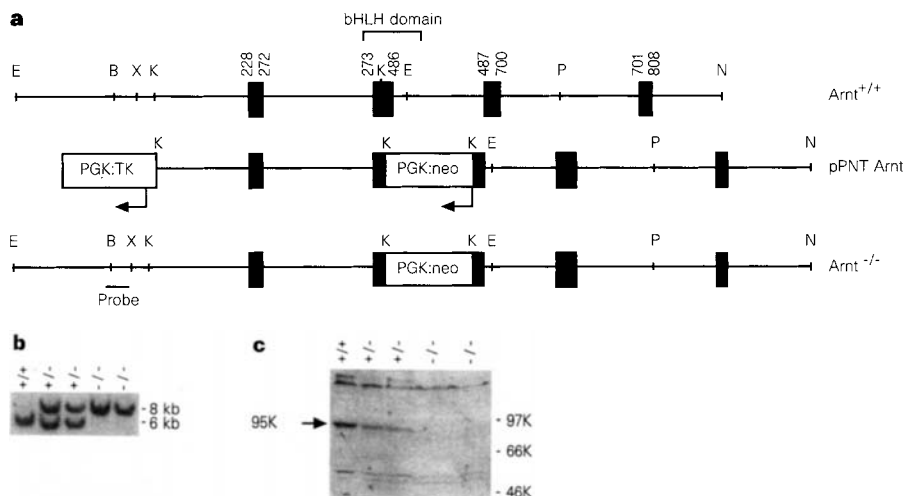


Figure 1 Targeted disruption of the *Arnt* gene. **a**, The *Arnt* targeting construct. Top, a partial restriction-endonuclease map of the murine *Arnt* gene; middle, the structure of the *Arnt* targeting construct pPNT *Arnt* containing the HSV-tk and neomycin(neo)-resistance genes, both under the control of the phosphoglycerate kinase (PGK) promoter; bottom, structure of the targeted *Arnt* allele that results in the insertion of the resistance gene within the bHLH domain, in the opposite orientation to *Arnt*. **b**, Southern blot analysis of the *Arnt* wild-type (+/+), heterozygous (+/-) and homozygous (-/-) mutant ES cell clones. The 6-kb *Eco*RI

band corresponds to the wild-type allele and the 8-kb band corresponds to the rearranged allele. **c**, Western blot of ARNT expression in wild-type, heterozygous and homozygous mutant ES cell clones. The anti-ARNT antibody detects a band migrating at *M_r* ~95K in the wild type and *Arnt*^{+/-} clone. Heterozygous clones express about half the amount of protein; *Arnt*^{-/-} ES cell clones do not express detectable amounts of ARNT. Upper and lower bands represent nonspecific immunoreaction.

increased production of PGK-1, ALDA, VEGF and GLUT-1 mRNAs over a 24-hour period, which was maximal at 16 hours (Fig. 2b, lanes 1–4). In contrast, two independently derived *Arnt*^{-/-} ES cell lines did not induce these genes during growth under hypoxic conditions (Fig. 2b, lanes 5–12). Results were similar for the glycolytic enzymes phosphofructokinase-L (PFK-L) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (data not shown). Two chemical agents that mimic physiological hypoxia, cobalt chloride and desferrioxamine, increased transcription of these genes in *Arnt*^{+/+} cells, but not in *Arnt*^{-/-} cells (Fig. 2c, d). Our results are in agreement with a report²¹ demonstrating a role for ARNT in regulating hypoxia-activated transcription in hepatocytes. We detected both ARNT and HIF-1 α mRNAs in untreated R1 ES cells (Fig. 2a, lane 1), the levels of which did not change in response to low O₂ tension (Fig. 2a, lanes 2–12).

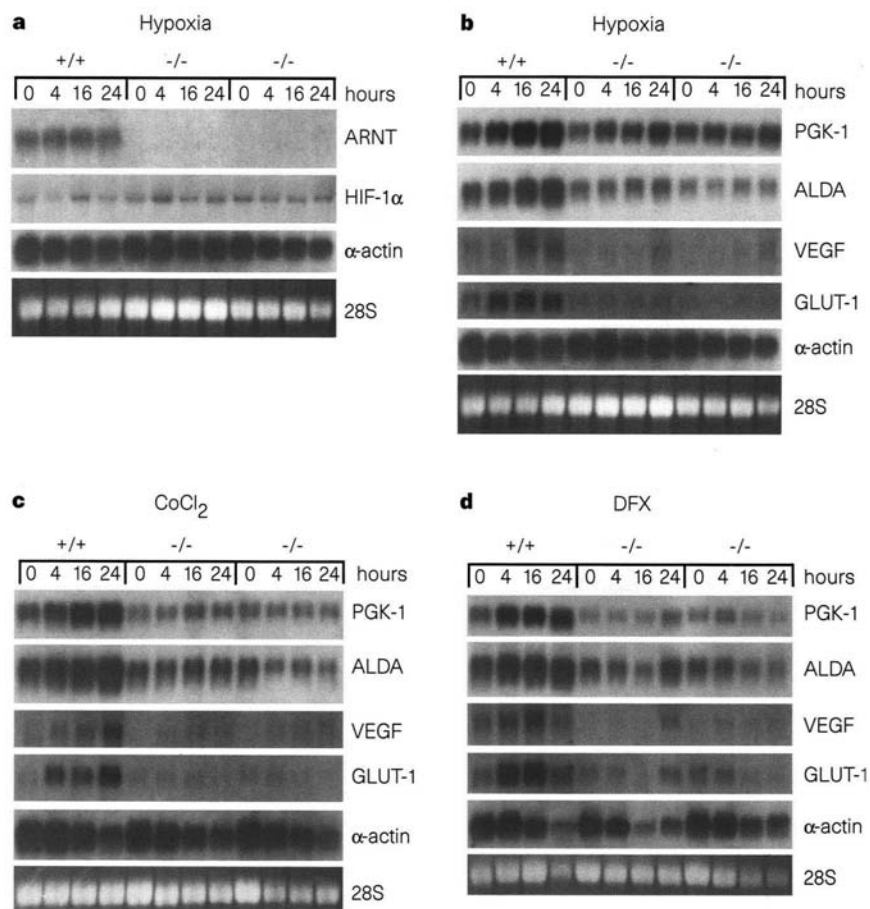
GLUT-1 and VEGF mRNA expression has been shown to increase in the C6 glioma cell line in response to both hypoxia and glucose deficiency^{22,23}. We therefore investigated the effect of glucose deprivation on GLUT-1, VEGF and other O₂-regulated genes in *Arnt*^{-/-} ES cells. *Arnt*^{+/+} ES cells grown in low-glucose medium for 36 h showed increased transcription of GLUT-1, VEGF and the glycolytic enzymes PGK-1 and ALDA, the kinetics and size of the effect being similar to those observed during culture under low O₂ tension (Fig. 3a, lanes 1–5). In contrast, *Arnt*^{-/-} ES cells failed to activate the expression of these genes in response to hypoglycaemia (Fig. 3a, lanes 6–15) and were less viable under these conditions, with both *Arnt*^{-/-} ES cell clones showing a 35–40% reduction in viability after 36 h and a 60% reduction after 72 h of glucose deprivation compared to *Arnt*^{+/+} control ES cells (Fig. 3b). *Arnt*^{+/+} cells had an intermediate phenotype and were 30% less viable than wild-type cells after 72 h (Fig. 3b). Thus, the inability of mutant cells to respond to the stress of glucose deprivation by stimulating the expression of genes such as GLUT-1 is correlated with decreased

survival. Our results provide evidence for the requirement for ARNT in the transcriptional response to glucose deficiency and indicate that the oxygen- and glucose-deprivation signalling pathways must converge to activate a common group of genes in a process that is mediated by ARNT.

To assess the role of ARNT *in vivo*, we generated *Arnt*^{-/-} mice from several independently derived *Arnt*^{+/+} ES cell clones. No *Arnt*^{-/-} mice were obtained from 104 live progeny of *Arnt*^{+/+} crosses, indicating that *Arnt*^{-/-} embryos die *in utero*. To determine the time of death resulting from the *Arnt* mutation, embryonic day (E) 8.5–16.5 embryos were examined. No viable *Arnt*^{-/-} embryos were found beyond E10.5. Most E9.5–10 *Arnt*^{-/-} embryos were readily distinguished from *Arnt*^{+/+} and *Arnt*^{+/-} littermates by the absence of blood-filled vessels in the yolk sac (Fig. 4a). Microscopic analysis showed that many E9.5 *Arnt*^{-/-} yolk sacs contained apparently normal blood islands but lacked any vasculature. Other *Arnt*^{-/-} yolk sacs lacked large vitelline vessels but contained enlarged capillaries instead (Fig. 4c–f). There were fewer capillaries and they appeared to be fused with one another, generating an irregular vascular plexus. A similar phenotype has been observed in both *Vegf*^{+/+} and tissue-factor-deprived (*TF*^{-/-}) yolk sacs^{14–16}. In addition, as for *TF*^{-/-} yolk-sac vessels¹⁶, the vessels in *Arnt*^{-/-} yolk sacs had fewer surrounding smooth-muscle cells, as indicated by using a monoclonal antibody against smooth-muscle actin (Fig. 4a, f). Histological staining revealed the presence of fused endothelial cells in mutant yolk sacs (data not shown). We propose that although vasculogenesis (the *de novo* formation of blood vessels) is intact, angiogenesis (the subsequent maturation of the vascular network) is abnormal in *Arnt*^{-/-} yolk sacs.

At E8.5, there were no morphological differences between *Arnt*^{+/+}, *Arnt*^{+/-} and *Arnt*^{-/-} embryos. E9.5 *Arnt*^{-/-} embryos, however, when dissected free of the yolk sac, were smaller and generally wasted, and development was delayed (Fig. 4b). Although

Figure 2 Northern analysis of RNA isolated from *Arnt*^{-/-} and *Arnt*^{+/+} ES cells cultured in 1.5% O₂ (**a** and **b**), 75 μ M CoCl₂ (**c**) or 130 μ M desferrioxamine (DFX) (**d**) for the indicated times. 10 μ g total RNA were loaded per lane. Equal loading and integrity of the RNA samples were confirmed by ethidium bromide staining of 28S rRNA in the gel before transfer to a nitrocellulose membrane. As expected, α -actin mRNA decreased during the same 24 hours, particularly in cells treated with cobalt or desferrioxamine.



developmentally stunted, individual organ systems in *Arnt*^{-/-} embryos appeared morphologically normal (Fig. 5a, b). Despite the extraembryonic vascularization being compromised, the heart and some of the embryonic vasculature, notably the dorsal aorta and intersomitic vessels, also appeared to be normal (Fig. 5a–d). This is in contrast to *Vegf*^{+/-} animals, which had generally defective vasculature and incorrect cardiac development^{14,15}. However, *Arnt*^{-/-} embryos had fewer blood vessels in solid tissues such as the branchial arch, as shown by anti-CD34 staining (Fig. 5c–f). We analysed *Arnt*^{+/-} and *Arnt*^{-/-} animals for VEGF expression by using *in situ* hybridization (Fig. 5). *Arnt*^{-/-} mice were found to have less VEGF mRNA in the yolk sac and in most of the embryo proper (Fig. 5i, j), but had high levels in the primitive gut. ARNT-independent expression of VEGF in this tissue may ensure development of the early heart. A homologue of ARNT, ARNT2, has been shown by *in situ* hybridization to be expressed predominantly in the neural tube of the mouse embryo²⁴. Both *Arnt*^{+/-} and *Arnt*^{-/-} embryos expressed ARNT2 (data not shown) and VEGF mRNA in the neural tube (Fig. 5i, j). Given the proximity of the dorsal aorta and intersomitic vessels to this neural tissue, it is likely that ARNT2-

driven VEGF production compensates here in *Arnt*^{-/-} animals. The vascularization of the remainder of the embryo and yolk sac, however, remains dependent on ARNT-mediated VEGF expression.

In conclusion, we have shown that ARNT plays a critical role in the transcriptional response to hypoxia and hypoglycaemia. *Arnt*^{-/-} embryos exhibit defective yolk-sac angiogenesis and abnormal development of the vitello-embryonic circulation, leading to a lethal fetal wasting owing to the demands on the circulation at and beyond E9.5. Capillary development in solid tissues in the embryo is also severely compromised. We propose that during organogenesis local hypoxic and hypoglycaemic conditions stimulate ARNT transcriptional activity, causing VEGF and tissue factor production and blood vessel development. Tissue factor increases in response to hypoxia²⁵, but this may be an indirect result of increased VEGF in these hypoxic tissues^{26,27}. It is notable here that mutations in the *Drosophila trachealess (trh)* gene, another bHLH-PAS protein, cause defects in the oxygen-delivery system (the trachea), suggesting that bHLH-PAS proteins may have an evolutionarily conserved role in regulating O₂ delivery^{28,29}. Our *Arnt*^{-/-} mice indicate that hypoxia and hypoglycaemia play a significant role in

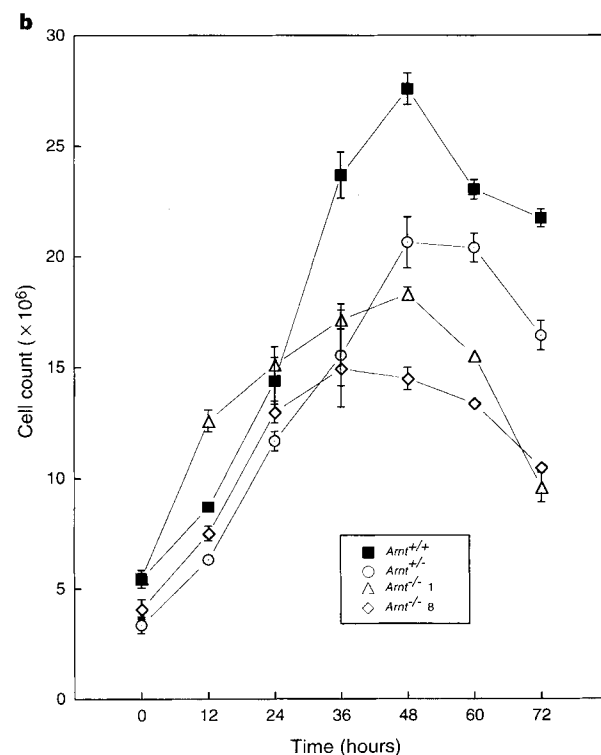
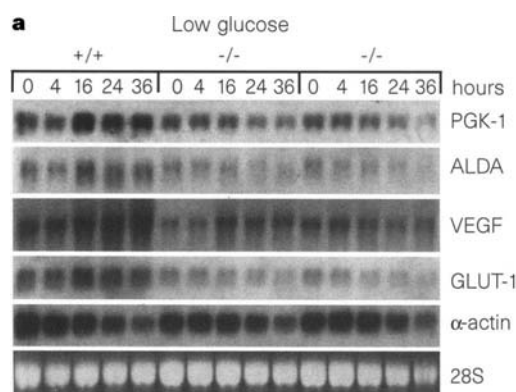
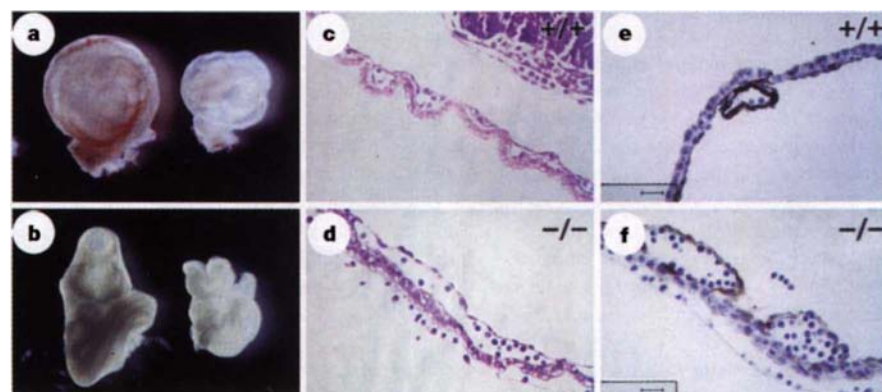


Figure 3 a, Northern analysis of RNA isolated from *Arnt*^{+/-} and *Arnt*^{-/-} ES cells cultured in low-glucose medium; 10 µg total RNA were loaded per lane. **b**, Viability of *Arnt*^{+/-}, *Arnt*^{+/-}, and *Arnt*^{-/-} ES cells cultured in low-glucose medium as assayed by trypan blue exclusion.

Figure 4 Gross appearance of *Arnt*^{+/-} and *Arnt*^{-/-} yolk sacs (**a**) and embryos (**b**) at E9.5. Material from *Arnt*^{+/-} animals is shown on the left and from mutants on the right. Note the apparent absence of vessels in the yolk sac and the developmental delay of the *Arnt*^{-/-} animals. **c**, **d**, Yolk-sac sections, showing a network of small vessels within the *Arnt*^{+/-} yolk sac compared to an abnormal, enlarged vessel in the *Arnt*^{-/-} yolk sac. **e**, **f**, Anti-smooth muscle actin immunostaining of yolk sacs, revealing the reduced number of smooth cells in *Arnt*^{-/-} vitelline vessels. Original magnification in **c**–**f**, 40 ×; scale bars, 50 µm.



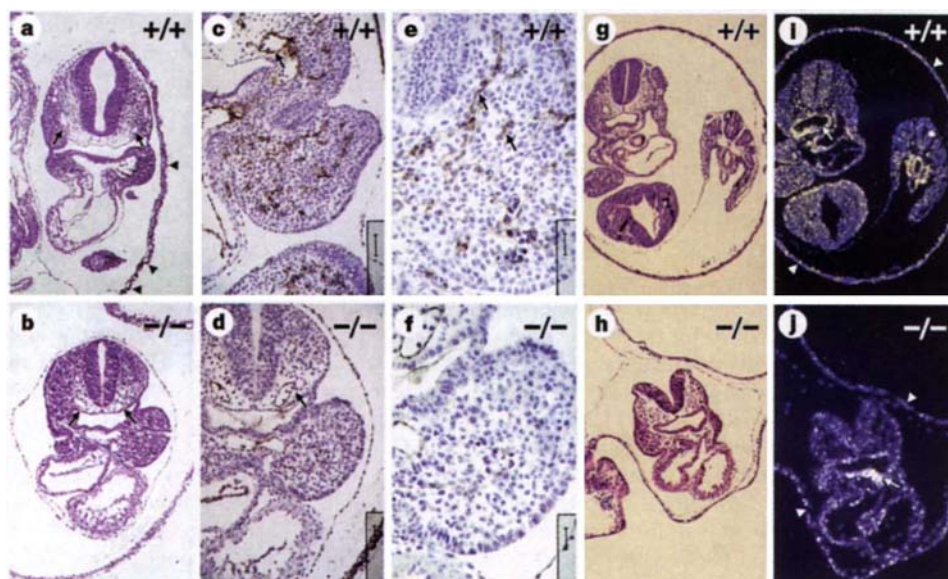


Figure 5 Haematoxylin- and eosin-stained *Arnt*^{+/+} (a) and *Arnt*^{-/-} (b) sections showing the presence of the dorsal aorta (arrows) and normal heart in both animals. Note the absence of normal vessels in the mutant yolk sac compared to wild type (arrowheads). c-f, Immunostaining of *Arnt*^{+/+} (c, e) and *Arnt*^{-/-} (d, f) sections with anti-CD34 antibodies, revealing a decrease in endothelial-cell-lined capillaries in the branchial arch of mutant embryos. Magnification in c, d: 20 ×; arrows indicate the dorsal aorta. Magnification in e, f, 40 ×; arrows indicate the

branchial arch blood vessels. Scale bars in c and d, 100 μm; in e, f, 50 μm. VEGF mRNA *in situ* hybridization in *Arnt*^{+/+} (i) and *Arnt*^{-/-} (j) embryos. Adjacent sections of haematoxylin-and-eosin-stained *Arnt*^{+/+} (g) and *Arnt*^{-/-} (h) animals are shown for comparison. Magnification in g and i, 10 ×; h and j, 20 ×. Note the high level of VEGF mRNA in the gut of both animals (arrows) and paucity of VEGF mRNA in the *Arnt*^{-/-} yolk sac (arrowheads) and additional embryonic structures.

regulating angiogenesis and vasoformation during mammalian embryonic development. □

Methods

Generation and characterization of *Arnt*^{-/-} ES cells. A P1 clone containing the murine *Arnt* gene was obtained from Genome Systems (St Louis, Missouri). The targeting construct was generated by insertion of a 2.7-kb *KpnI* fragment containing the 5' end of the bHLH exon of the *Arnt* gene into the *KpnI* site of pPNT³⁰, followed by ligation of a 5.2-kb *KpnI*/*NotI* fragment containing the 3' end of the same bHLH domain. The resulting construct was linearized by *NotI* digestion and electroporated into R1 ES cells. Homologous recombinants selected by G418 (250 μg ml⁻¹) and Gancyclovir (1 mM) were screened by Southern blotting for presence of a recombinant 8-kb *EcoRI* band, in addition to the endogenous 6-kb fragment, using a 300-bp *BamHI*/*XbaI* fragment from the *Arnt* gene. Heterozygous clones were passaged onto gelatin-coated plates and subjected to 3 mg ml⁻¹ G418 selection to obtain *Arnt*^{-/-} cells. Protein was extracted from these clones and separated by SDS-PAGE, transferred to nitrocellulose, and probed with a polyclonal antibody generated against the C terminus of the ARNT protein.

Cell culture. Gelatin-adapted R1 ES cells were grown to ~70% confluence in medium consisting of high-glucose DMEM (Gibco-BRL) supplemented with 15% fetal calf serum (Hyclone), penicillin-streptomycin (Gibco-BRL), MEM non-essential amino acids (Gibco-BRL) and leukaemia inhibitory factor. Hypoxic conditions were established by culturing cells in an atmosphere of 1.5% O₂/5% CO₂/93.5% N₂ generated by a gas mixer, or by culturing in the presence of 75 μM CoCl₂ (Sigma) or 130 mM desferrioxamine (Sigma). For low-glucose challenge, glucose-free DMEM (Gibco-BRL) was used.

Northern analysis. Cells cultured for the indicated times were washed once with PBS (Gibco-BRL) and RNA was extracted with TRIzol reagent (Gibco-BRL) according to the manufacturer's instructions. Full-length human GLUT-1 cDNA was provided by C. Burant. Other probes included a 249-bp HIF-1α cDNA fragment, a 0.9-kb murine α-actin cDNA, and RT-PCR products generated with 3'-UTR-specific primers for PGK-1, ALDA, VEGF, ARNT and ARNT2.

Immunohistochemistry. 4-μm sections of each sample were incubated overnight at 4 °C with monoclonal antibodies generated against CD34 (Phar-

Mingen) and smooth-muscle α-actin (Sigma). Immunohistochemical staining was done on a Ventana Gen System (Ventana Medical Systems) which uses an indirect streptavidin-biotin system conjugated with horseradish peroxidase.

Received 8 November 1996; accepted 27 January 1997.

- Burbach, K. M., Poland, A. & Bradfield, C. A. Cloning of the Ah-receptor cDNA reveals a distinctive ligand-activated transcription factor. *Proc. Natl Acad. Sci. USA* **89**, 8185-8189 (1992).
- Wang, G. L., Jiang, B.-H., Rue, E. A. & Semenza, G. L. Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O₂ tension. *Proc. Natl Acad. Sci. USA* **92**, 5510-5514 (1995).
- Ema, M. *et al.* cDNA cloning of a murine homologue of *Drosophila* single-minded, its mRNA expression in mouse development, and chromosome localization. *Biochem. Biophys. Res. Commun.* **218**, 588-594 (1996).
- Hoffman, E. C. *et al.* Cloning of a factor required for activity of the Ah (dioxin) receptor. *Science* **252**, 954-958 (1991).
- Swanson, H. L., Chan, W. K. & Bradfield, C. A. DNA binding specificities and pairing rules of Ah receptor, ARNT, and SIM proteins. *J. Biol. Chem.* **270**, 26292-26302 (1995).
- Sogawa, K. *et al.* Possible function of Ah receptor nuclear translocator (Arnt) homodimer in transcriptional regulation. *Proc. Natl Acad. Sci. USA* **92**, 1936-1940 (1995).
- Nebert, D. W. & Gonzales, F. J. P450 genes: structure, evolution, and regulation. *Annu. Rev. Biochem.* **56**, 945-993 (1987).
- Telakowski-Hopkins, C., King, R. & Pickett, C. Glutathione S-transferase Ya subunit gene: identification of regulatory elements required for basal level and inducible expression. *Proc. Natl Acad. Sci. USA* **85**, 1000-1004 (1988).
- Hankinson, O. The aryl hydrocarbon receptor complex. *Annu. Rev. Pharmacol. Toxicol.* **35**, 307-340 (1995).
- Wang, G. L. & Semenza, G. L. General involvement of hypoxia-inducible factor in transcriptional response to hypoxia. *Proc. Natl Acad. Sci. USA* **90**, 4304-4308 (1993).
- Forsythe, J. A. *et al.* Activation of vascular endothelial growth factor gene transcription by hypoxia-inducible factor 1. *Mol. Cell. Biol.* **16**, 4604-4613 (1995).
- Wang, G. L. & Semenza, G. L. Characterization of hypoxia-inducible factor 1 and regulation of DNA binding activity by hypoxia. *J. Biol. Chem.* **268**, 21513-21518 (1993).
- Bunn, H. F. & Poyton, R. O. Oxygen sensing and molecular adaptation to hypoxia. *Physiol. Rev.* **76**, 839-885 (1996).
- Carmeliet, P. *et al.* Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele. *Nature* **380**, 435-439 (1996).
- Ferrara, N. *et al.* Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene. *Nature* **380**, 439-442 (1996).
- Carmeliet, P. *et al.* Role of tissue factor in embryonic blood vessel development. *Nature* **383**, 73-75 (1996).
- Mortensen, R., Conner, D., Chao, S., Geisterfer-Lowrance, A. & Seidman, J. Production of homozygous mutant ES cells with a single targeting construct. *Mol. Cell. Biol.* **12**, 2391-2394 (1992).
- Shweiki, D., Itin, A., Soffer, D. & Keshet, E. Vascular endothelial growth factor induced by hypoxia may mediate hypoxia-initiated angiogenesis. *Nature* **359**, 843-845 (1992).
- Firth, J. D., Ebert, B. L., Pugh, C. W. & Ratcliffe, P. J. Oxygen-regulated control elements in the phosphoglycerate kinase 1 and lactate dehydrogenase A genes: Similarities with the erythropoietin 3' enhancer. *Proc. Natl Acad. Sci. USA* **91**, 6496-6500 (1994).
- Semenza, G. L., Roth, P. H., Fang, H.-M. & Wang, G. L. Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1. *J. Biol. Chem.* **269**, 23757-23769 (1994).

21. Wood, S. M., Gleadle, J. M., Pugh, C. W., Hankinson, O. & Ratcliffe, P. J. The role of aryl hydrocarbon receptor nuclear translocator (ARNT) in hypoxia induction of gene expression. *J. Biol. Chem.* **271**, 15117–15123 (1996).
22. Shweiki, D., Neeman, M., Itin, A. & Keshet, E. Induction of vascular endothelial growth factor expression by hypoxia and by glucose deficiency in multicell spheroids: Implications for tumor angiogenesis. *Proc. Natl Acad. Sci. USA* **92**, 768–772 (1995).
23. Stein, I., Neeman, M., Shweiki, D., Itin, A. & Keshet, E. Stabilization of vascular endothelial growth factor mRNA by hypoxia and hypoglycemia and coregulation with other ischemia-induced genes. *Mol. Cell. Biol.* **15**, 5363–5368 (1995).
24. Hirose, K. *et al.* cDNA cloning and tissue-specific expression of a novel basic helix-loop-helix/PAS factor (Arnt2) with close sequence similarity to the Aryl hydrocarbon receptor nuclear translocator (Arnt). *Mol. Cell. Biol.* **16**, 1706–1713 (1996).
25. Compeau, C. G. *et al.* *In situ* ischemia and hypoxia enhance alveolar macrophage tissue factor expression. *Am. J. Respir. Cell Mol. Biol.* **11**, 446–455 (1994).
26. Clauss, M. *et al.* Vascular permeability factor: A tumor-derived polypeptide that induces endothelial cell and monocyte procoagulant activity, and promotes monocyte migration. *J. Exp. Med.* **172**, 1535–1545 (1990).
27. Zhang, Y. *et al.* Tissue factor controls the balance of angiogenic and antiangiogenic properties of tumor cells in mice. *J. Clin. Invest.* **94**, 1320–1327 (1994).
28. Wilk, R., Weizman, I. & Shilo, B.-Z. *tracheless* encodes a bHLH-PAS protein that is an inducer of tracheal cell fates in *Drosophila*. *Genes Dev.* **10**, 93–102 (1996).
29. Isaac, D. D. & Andrew, D. J. Tubulogenesis in *Drosophila*: a requirement for the tracheless gene product. *Genes Dev.* **10**, 103–117 (1996).
30. Tybulewicz, V., Crawford, C., Jackson, P., Bronson, R. & Mulligan, R. Neonatal lethality and lymphopenia in mice with a homozygous disruption of the *c-abl* proto-oncogene. *Cell* **65**, 1153–1163 (1991).

Acknowledgements. We thank C. Clendenin, K. Sigrist and M. M. Lu for technical assistance; N. Chandel for help with hypoxic cell culture; J. Leiden, M. Parmacek and B. Keith for critically reviewing the manuscript; C. Small for secretarial assistance; and L. Gottschalk for preparing the illustrations. E.M. is a Medical Scientist Trainee. M.C.S. is an investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to M.C.S. (e-mail: csimon@medicine.bsd.uchicago.edu).

Sensitization of diabetic and obese mice to insulin by retinoid X receptor agonists

Ranjan Mukherjee*, Peter J. A. Davies†, Diane L. Crombie‡, Eric D. Bischoff‡, Rosemary M. Cesario‡, Lily Jow*, Lawrence G. Hamann§, Marcus F. Boehm||, Carl E. Mondon†, Alex M. Nadzan||, James R. Paterniti Jr* & Richard A. Heyman‡

Departments of * Retinoid Research, * Cardiovascular Research, || Retinoid Chemistry, and § Endocrine Chemistry, Ligand Pharmaceuticals, 10255 Science Center Drive, San Diego, California 92121, USA
† Department of Pharmacology & Medicine, University of Texas Medical School, Health Science Center at Houston, Houston, Texas 77225, USA
¶ METABOLEX Inc., 3876 Bay Center Place, Hayward, California 94545, USA

Retinoic acid receptors (RAR), thyroid hormone receptors (TR), peroxisome proliferator activated receptors (PPARs) and the orphan receptor, LXR, bind preferentially to DNA as heterodimers with a common partner, retinoid X receptor (RXR), to regulate transcription^{1–6}. We investigated whether RXR-selective agonists replicate the activity of ligands for several of these receptors? We demonstrate here that RXR-selective ligands (referred to as rexinoids) function as RXR heterodimer-selective agonists, activating RXR:PPAR γ and RXR:LXR dimers but not RXR:RAR or RXR:TR heterodimers. Because PPAR γ is a target for antidiabetic agents, we investigated whether RXR ligands could alter insulin and glucose signalling. In mouse models of non-insulin-dependent diabetes mellitus (NIDDM) and obesity, RXR agonists function as insulin sensitizers and can decrease hyperglycaemia, hypertriglyceridaemia and hyperinsulinaemia. This antidiabetic activity can be further enhanced by combination treatment with PPAR γ agonists, such as thiazolidinediones. These data suggest that the RXR:PPAR γ heterodimer is a single-function complex serving as a molecular target for treatment of insulin resistance. Activation of the RXR:PPAR γ dimer with rexinoids may provide a new and effective treatment for NIDDM.

We have recently identified the synthetic RXR-selective ligands,

LG100268⁷ and LGD1069⁸, that specifically bind to RXRs and activate RXR homodimer-dependent transcription^{7–9}. Because numerous non-steroid receptors function as heterodimers with RXR, we used a co-transfection assay to examine RXR responsiveness in the context of different RXR dimers (Fig. 1a). Cells transfected with RXR:PPAR γ or RXR:LXR respond to PPAR γ ligands (such as the thiazolidinedione (TZD)) BRL49653^{10–12} or LXR activators (such as 22(R)-OH-cholesterol)¹³. Furthermore, these heterodimers are also activated by RXR ligands, such as

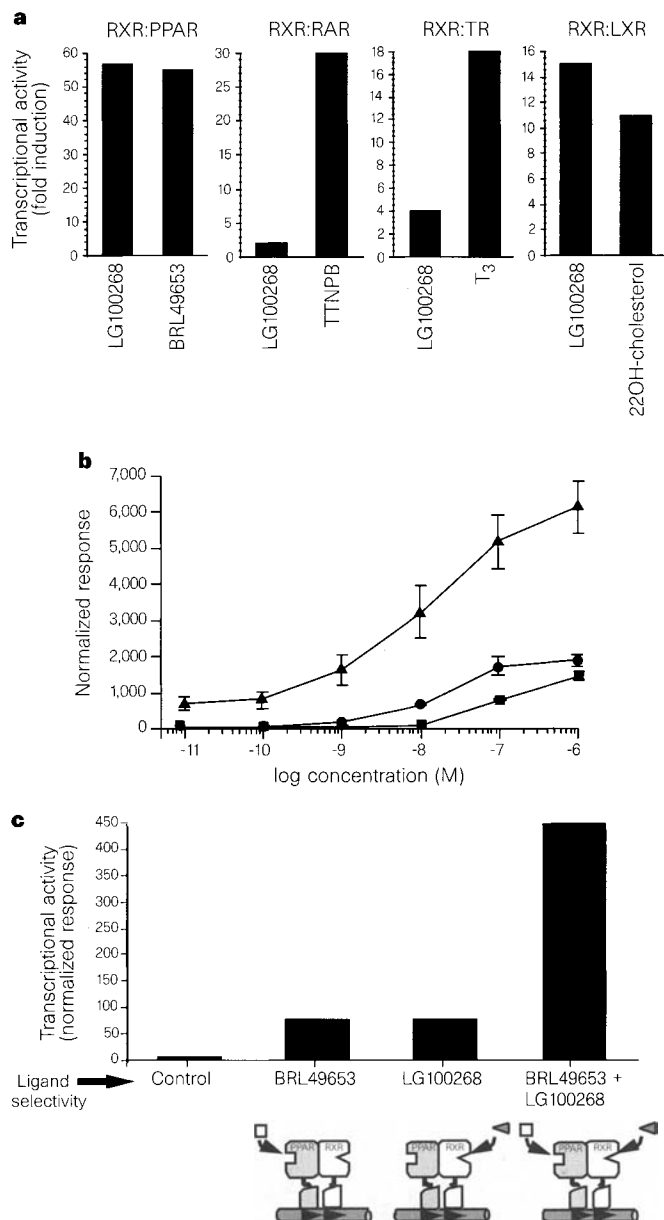


Figure 1 Activation of RXR heterodimers by LG100268 is dimer specific. **a**, Transient transfection analysis of RXR:PPAR γ , RXR:RAR, RXR:TR and RXR:LXR heterodimers. CV-1 cells cotransfected with RXR along with various receptors and their cognate luciferase reporters (see Methods) were treated with either LG100268 (1 μ M) for all dimer combinations or BRL49653 (1 μ M) for RXR:PPAR γ , TNPB (100 nM) for RXR:RAR, T₃ (100 nM) for RXR:TR and 22(R)-OH-cholesterol (5 μ M) for RXR:LXR dimers. **b**, LG100268 and BRL49653 dose-response curves of RXR:PPAR γ heterodimer. CV-1 cells were cotransfected with RXR and PPAR γ plasmids along with PPRE luciferase reporter. Increasing concentrations of LG100268 (●) or BRL49653 (■) alone or increasing amounts of LG100268 with a fixed concentration of BRL49653 (100 nM) (▲). **c**, LG100268 and BRL49653 synergistically transactivate RXR:PPAR γ heterodimer. CV-1 cells cotransfected as above; LG100268 and BRL49653 were added at 3.0 nM and 30 nM, respectively.

LG100268. The ligand-dependent activation of RXR:PPAR γ and RXR:LXR is in sharp contrast to RXR:TR and RXR:RAR which LG100268 does not activate, although these dimers respond to triiodothyronine (T₃) and the RAR-selective ligand, TTNPB¹⁴, respectively (Fig. 1a and refs 15, 16). Thus, RXR:PPAR γ and RXR:LXR have dual-ligand responsiveness, whereas RXR:RAR and RXR:TR do not. A detailed examination of RXR:PPAR γ activity demonstrates that LG100268 produces a concentration-dependent increase in transactivation with a half-maximal effective concentration (EC₅₀) of 20 nM and a maximal luciferase induction of 57-fold at 1.0 μ M (Fig. 1b). Under the same conditions that LG100268 activates RXR:PPAR γ , BRL49653 produces a concentration-dependent increase in transactivation with a maximal luciferase induction of 55-fold at 1.0 μ M (Fig. 1b). Furthermore, the magnitude of the LG100268 dose-response is significantly enhanced in the presence of a fixed concentration of BRL49653. At submaximal levels of LG100268 and BRL49653 (3.0 nM and 30 nM, respectively) there is a synergistic transcriptional response when both ligands are present (Fig. 1c). These results demonstrate that the RXR:PPAR γ

heterodimer: (1) responds to ligands that bind to either receptor subunit; (2) the maximal activation requires both ligands; and (3) the response with both ligands is synergistic.

Recent studies have shown that the ability of TZDs to bind and activate PPAR γ correlates with their ability to reduce hyperglycaemia in animal models of NIDDM and obesity¹⁷. The *ob/ob* and *db/db* strains of mice have a syndrome characterized by hyperglycaemia, hypertriglyceridaemia and hyperinsulinaemia^{18,19}; this syndrome is secondary to defects in leptin signalling pathways^{20,21}. In these animals, TZDs function as insulin sensitizers decreasing elevated serum glucose, triglycerides and insulin²². Because TZDs are thought to exert their antidiabetic activity via activation of the RXR:PPAR γ heterodimer, we speculated that RXR ligands might also alter glucose homeostasis in these animal models. To address this issue, we examined the effects of RXR-selective ligands on glucose, triglycerides and insulin concentrations in *ob/ob* and *db/db* mice and compared their activity to a TZD. At 53 days of age, the average fasting glucose levels of the *ob/ob* mice is 262 mg dl⁻¹ (Fig. 2b). During the subsequent 14 days, control *ob/ob* mice become

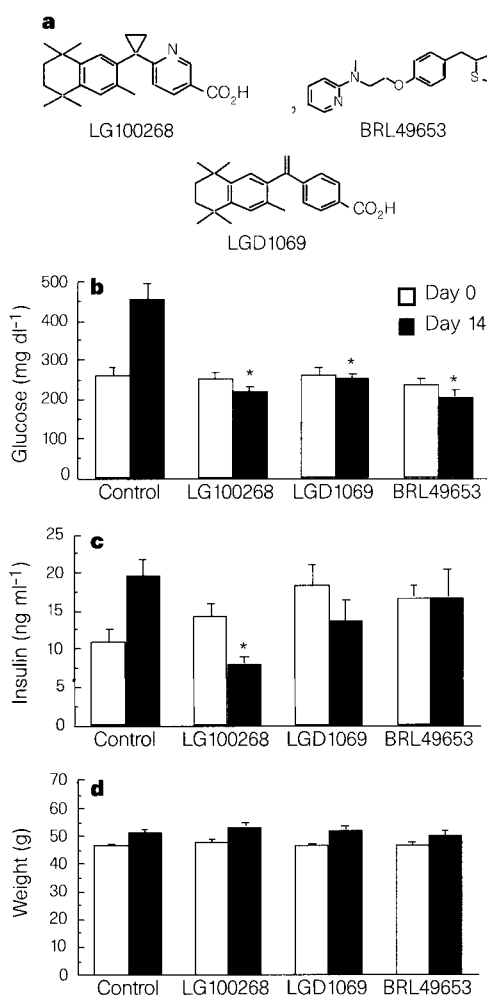


Figure 2 RXR agonists decrease hyperglycaemia and hyperinsulinaemia in *ob/ob* mice. **a**, Structures of LGD1069, LG100268 and BRL49653. **b**, Glucose, **c**, insulin and **d**, body weight were measured in *ob/ob* mice dosed with vehicle, LG100268 (20 mg kg⁻¹), LGD1069 (30 mg kg⁻¹) or BRL49653 (0.4 mg kg⁻¹). Mice (9 animals per group) were gavaged once daily for 14 days. Fasting plasma glucose and insulin levels were measured after a 3 h fast on day 0 (white bars) or on day 14 (black bars). Each point represents the mean value $\pm$ s.e.m. * Significantly different ($P < 0.01$) from the vehicle-treated control group as determined by Student's *t*-test. No effect on food intake was observed in the different treatment groups (data not shown).

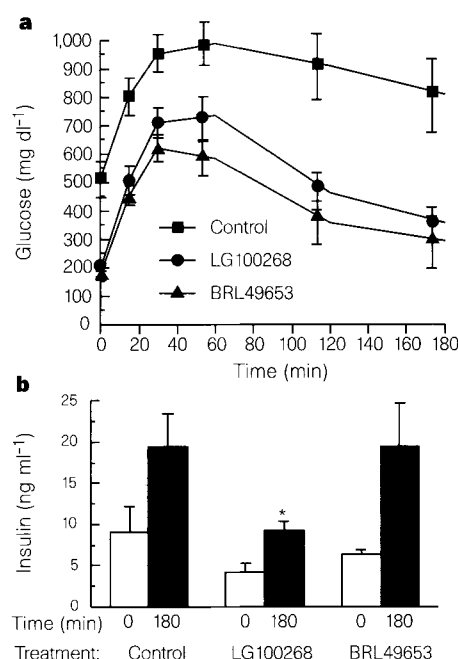


Figure 3 RXR agonist and PPAR γ agonist alter glucose disposal in glucose tolerance test. **a**, Glucose and **b**, insulin concentrations in *ob/ob* mice were monitored during a glucose tolerance test. Mice were treated with vehicle, LG100268 (20 mg kg⁻¹) and BRL49653 (3 mg kg⁻¹) as in Fig. 2 followed by an intraperitoneal injection of a 5.0% glucose solution (1.125 g kg⁻¹). Plasma glucose and insulin were measured at $t = 0, 15, 30, 60, 120$ and 180 min. In **b**, white bars are insulin concentrations at $t = 0$ and black bars are insulin concentrations at $t = 180$ min. * Significantly different ($P < 0.05$) from vehicle-treated control group at 180 min as determined by Student's *t*-test.

Table 1 Glucose tolerance test (GTT) and insulin resistance (IR) index in *ob/ob* mice treated with RXR and PPAR γ ligands

Group	n	GTT area (cm ²)		IR index
		Glucose	Insulin	Glucose $\times$ insulin
Vehicle	(8)	26.9 $\pm$ 3.3	34.5 $\pm$ 6.6	8.03 $\pm$.85
BRL49653	(8)	13.7 $\pm$ 1.9†	28.2 $\pm$ 5.1	4.11 $\pm$ 1.00*
LG100268	(9)	15.3 $\pm$ 1.2†	14.7 $\pm$ 2.1*	2.21 $\pm$.28*

Values are mean $\pm$ s.e.m.; n = numbers of animals per group. Area of glucose and insulin curves was calculated by multiplying the cumulative mean height of the glucose values (1 mg ml⁻¹ = 1 cm) and insulin values (1 ng ml⁻¹ = 1 cm), respectively, by time (60 min = 1 cm). IR index was calculated from the product of the area of glucose and insulin $\times 10^{-2}$ as described in ref. 29. Comparisons between vehicle and test compounds were made using Student's *t* test. † $P < 0.01$; * $P < 0.02$.

progressively hyperglycaemic and their fasting glucose levels increase up to 450 mg dl^{-1} . Daily treatment with LG100268 (20 mg kg^{-1}) (structures shown in Fig. 2a) for 14 days not only prevents the time-dependent increase in glucose concentrations, but results in a slight fall below the initial values. After 14 days of LG100268 treatment, fasting glucose concentrations are 52% of control animals. Another RXR-selective compound, LGD1069 (Targretin, Fig. 2a)⁸, which is presently undergoing phase I/II human clinical trials for cancer therapy, decreases fasting glucose levels to 55% of control levels. For comparison, treatment with BRL49653 (0.4 mg kg^{-1}) results in a similar decrease in glucose levels to 45% of control levels. The control *ob/ob* mice are severely hyperinsulinaemic with fasting insulin concentrations that range from $12\text{--}18 \text{ ng ml}^{-1}$ (Fig. 2c). LG100268 and LGD1069 treatment decreases insulin levels to 41% and 70%, respectively of controls, whereas BRL decreases the insulin levels to 85% of control values. The decrease in glucose and insulin was not a secondary response to changes in body weight because all animals gained weight at comparable rates (Fig. 2d). Thus, RXR compounds produce a coordinated decrease in both glucose and insulin concentrations suggesting an overall improvement in glucose homeostasis.

We conducted a glucose tolerance test (GTT) to determine whether RXR agonists reduce insulin resistance by increasing glucose disposal. Mice were treated with vehicle, LG100268 or BRL49653 for 14 days and then given a bolus glucose injection. Glucose and insulin levels were then measured during the subsequent 180 min. In control animals, there is an elevated and prolonged response to the glucose such that at 3 h after injection, glucose levels are 65% above basal (Fig. 3a). LG100268 and BRL49653 treatments both have lower glucose levels at $t = 0$ min, glucose peaks at 60 min and approaches basal levels by 180 min. To compare glucose disposal between the treatment groups, the area of the glucose curve (GTT glucose area, Table 1) was determined. The values for the groups treated with control, LG100268 and BRL49653 are 26.9, 15.3 and 13.7, respectively, demonstrating enhanced glucose disposal in the LG100268- and TZD-treated animals. What is strikingly different is the insulin response to the different treatments. LG100268-treated animals have a lower basal level of insulin at both initiation ($t = 0$ min) and conclusion ($t = 180$ min) of the GTT (Fig. 3b). Insulin levels in the BRL49653-treated group are lower than control at initiation but remain elevated and are similar to control levels at 180 min. These differences are reflected in the GTT insulin areas (Table 1) of 34.5, 14.7 and 28.2 for the animals treated with control, LG100268 and BRL49653, respectively. The insulin resistance index (IR index) (Table 1) was determined to compare the resistance to insulin action between the groups and is derived from the product of the glucose and insulin curves. A large IR index reflects elevated insulin resistance and thus insulin is less effective in lowering plasma glucose levels. LG100268 results in a fourfold lower IR index, whereas BRL49653 only halves the IR index compared to the control group. These data suggest that rexinoids enhance both the kinetics of glucose disposal and increase sensitivity to endogenous insulin.

The *db/db* mice are an alternative genetic model of NIDDM. These animals are extremely hyperglycaemic, with fasting glucose concentrations greater than 500 mg dl^{-1} by 7 weeks of age (Fig. 4a). Similar to *ob/ob* mice, the diabetic status of these animals progresses with time as shown by a continuous increase in fasting serum glucose (Fig. 4a) as well as triglyceride concentrations (Fig. 4b). LG100268 (20 mg kg^{-1}) lowers fasting glucose to a new steady-state level within 2–3 days which remains reduced throughout the duration of study. After 14 days of treatment, glucose levels are reduced to 68% of control values. Significant decreases in fasting glucose levels occur at LG100268 doses as low as 1 mg kg^{-1} (data not shown). BRL49653 (1.0 mg kg^{-1}) alone also reduces the fasting glucose levels. To determine if the response to RXR agonists could be enhanced by TZDs, animals were treated with LG100268 in the

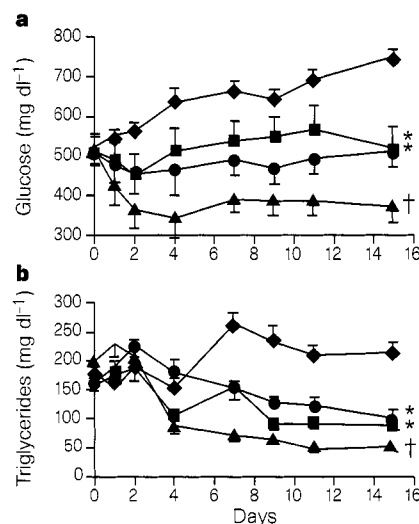


Figure 4 Thiazolidinediones increase the efficacy of RXR agonist in lowering fasting glucose and triglyceride levels in *db/db* mice. **a**, BRL49653 enhances the efficacy of LG100268 in decreasing fasting hyperglycaemia. **b**, BRL49653 enhances efficacy of LG100268 in decreasing fasting hypertriglyceridaemia. *Db/db* mice were dosed with vehicle (◆), LG100268 (●) at 20 mg kg^{-1} , BRL49653 (■) at 1 mg kg^{-1} or combination of LG100268 and BRL49653 (▲) at 20 mg kg^{-1} and 1 mg kg^{-1} , respectively, for 15 days as described in Fig. 3 (9 animals per treatment group). Animals were fasted and bled as above on the days shown. Each point represents the mean value $\pm$ s.e.m. *Significantly different ($P < 0.01$) from the vehicle-treated control group as determined by Student's *t*-test. †Significantly different ($P < 0.001$) in comparison to single agent treatment with LG100268 or BRL49653. No effect on body weight or food intake was observed in the different treatment groups (data not shown).

presence or absence of BRL49653. Combination treatment with LG100268 and BRL49653 proved to be much more efficacious than either compound alone resulting in glucose levels at 48% of the control group. Hypertriglyceridaemia is also a characteristic trait of insulin resistance in rodent models and is an important risk factor for human cardiovascular disease. In *db/db* mice, LG100268 or BRL49653, as single agents, markedly reduce serum triglycerides by 54% after 14 days of treatment (Fig. 4b). In addition, cotreatment with TZD and LG100268 result in an even greater decrease. Thus, RXR agonists reduce glucose and triglyceride levels in *db/db* mice and this response is enhanced by TZDs.

These studies demonstrate that RXR agonists function as insulin sensitizers, markedly decreasing serum glucose, triglycerides and insulin in two animal models of insulin resistance. The combination of RXR and PPAR γ ligands has additive or synergistic effects in stimulating transcription and reducing the degree of hyperglycaemia and hypertriglyceridaemia in animal models of NIDDM. The biological activity of RXR ligands is different from that of retinoids that activate the RARs, such as all-*trans* retinoic acid. Therefore, it is appropriate that RXR activators be considered a distinct class of pharmacological agents and could be referred to as 'rexinoids' to distinguish their activity from 'retinoids'; a term introduced by Sporn and colleagues 20 years ago to describe vitamin A and its synthetic analogues²³. Rexinoids appear to have pharmacological properties comparable with drugs with proven efficacy in humans for treatment of NIDDM²⁴. Our results indicate that activation of the RXR:PPAR γ functional unit is the common molecular target of thiazolidinedione and rexinoid therapy leading to regulation of gene expression and an improvement of glucose homeostasis. These RXR agonists appear to offer a new approach for the treatment of insulin resistance. Furthermore, it is possible that specific classes of RXR dimer-selective ligands can be identified that activate different RXR heterodimers. This may contribute to the treatment of certain

endocrine disorders by altering transcriptional activity of different RXR-dependent receptor complexes. □

Methods

Cell culture and transfection. CV-1 cells were cultured in DMEM with 10% fetal calf serum (FCS) and transfected as previously described⁹. To form heterodimers, RXR α ¹⁴ was cotransfected along with an expression vector for either PPAR γ ²⁵, RAR α ²⁶, TR β ⁶ or LXR β . RXR:PPAR γ activity was tested on the PPRE-tk-LUC reporter as described⁵; RXR:RAR α and RXR:TR activity were tested on a DR-5-tk-LUC and DR4-tk-LUC reporter²⁶, respectively. LXR:RXR was tested on a LXRE-tk-LUC as described⁶. Transcriptional activity (fold induction) was determined on the basis of the luciferase value divided by the β -gal activity as described in ref. 27. Each experiment was repeated at least three times.

In vivo mouse models of diabetic and obese mice. Female obese C57BL/6J-Lep^{ob}(4) mice (53 days old at commencement of study) were dosed with vehicle (0.9% carboxymethyl cellulose, 9.95% polyethylene glycol and 0.05% Tween 80), or drugs (20 mg kg⁻¹) by gavage once daily for 14 days. The gavage volume was 0.6 ml per 42 g. Blood was drawn after a 3-h fast on indicated days from the tip of the tail and plasma glucose and triglyceride levels were measured in duplicate by adoption of the glucose oxidase Trinder protocol²⁸ and Sigma triglyceride (GPO-Trinder) reagent. Plasma insulin was measured by Linco insulin immunoassay against rat insulin antibody (Linco Research, St Louis, Missouri). During the GTT, collection of blood was conducted from the tip of the tail in partially sedated animals (held in soft cloth towel) to avoid undue stress. Female diabetic C57BLKS/J-m+/+db mice (47 days old at commencement of study) were dosed as described above with vehicle, LG100268 at 20 mg kg⁻¹, BRL49653 at 1 mg kg⁻¹ or combination of LG100268 and BRL49653 at 20 mg kg⁻¹ and 1 mg kg⁻¹, respectively for 15 days. Animals were fasted and bled as above.

Received 21 October 1996; accepted 13 February 1997.

- Yu, V. C. *et al.* RXR β : a coregulator that enhances binding of retinoic acid, thyroid hormone, and vitamin D receptors to their cognate response elements. *Cell* **67**, 1251–1266 (1991).
- Zhang, X. K., Hoffman, B., Tran, P. B., Graupner, G. & Pfahl, M. Retinoid X receptor is an auxiliary protein for thyroid hormone and retinoic acid receptors. *Nature* **355**, 441–446 (1992).
- Leid, M. *et al.* Purification, cloning, and RXR identity of the HeLa cell factor with which RAR or TR heterodimerizes to bind target sequences efficiently. *Cell* **68**, 377–395 (1992).
- Kliwer, S. A., Umesono, K., Mangelsdorf, D. J. & Evans, R. M. Retinoid X receptor interacts with nuclear receptors in retinoic acid, thyroid hormone and vitamin D₃ signalling. *Nature* **355**, 446–449 (1992).
- Kliwer, S. A., Umesono, K., Noonan, D. J., Heyman, R. A. & Evans, R. M. Convergence of 9-*cis* retinoic acid and peroxisome proliferator signalling pathways through heterodimer formation of their receptors. *Nature* **358**, 771–774 (1992).
- Willy, P. J. *et al.* LXR, a nuclear receptor that defines a distinct retinoid response pathway. *Genes Dev.* **9**, 1033–1045 (1995).
- Boehm, M. F. *et al.* Design and synthesis of potent retinoid X receptor selective ligands that induce apoptosis in leukemia cells. *J. Med. Chem.* **38**, 3146–3155 (1995).
- Boehm, M. F. *et al.* Synthesis and structure-activity relationships of novel retinoid X receptor selective retinoids. *J. Med. Chem.* **37**, 2930–2941 (1994).
- Lala, D. *et al.* Activation of specific RXR heterodimers by an antagonist of RXR homodimers. *Nature* **383**, 450–453 (1996).
- Lehman, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPAR gamma). *J. Biol. Chem.* **270**, 12953–12956 (1995).
- Kliwer, S. A. *et al.* A prostaglandin J₂ metabolite binds peroxisome proliferator-activated receptor gamma and promotes adipocyte differentiation. *Cell* **83**, 813–819 (1995).
- Forman, B. M. *et al.* 15-Deoxy-delta 12,14-postaglandin J₂ is a ligand for the adipocyte determination factor PPAR gamma. *Cell* **83**, 803–812 (1995).
- Janowski, B. A., Willy, P. J., Devi, T. R., Falck, J. R. & Mangelsdorf, D. J. An oxysterol signalling pathway mediated by the nuclear receptor LXR α . *Nature* **383**, 728–731 (1996).
- Mangelsdorf, D. J., Ong, E. S., Dyck, J. A. & Evans, R. M. Nuclear receptor that identifies a novel retinoic acid response pathway. *Nature* **345**, 224–229 (1990).
- Kurokawa, R. *et al.* Regulation of retinoid signalling by receptor polarity and allosteric control of ligand binding. *Nature* **371**, 528–531 (1994).
- Forman, B. M., Umesono, K., Chen, J. & Evans, R. M. Unique response pathways are established by allosteric interactions among nuclear hormone receptors. *Cell* **81**, 541–550 (1995).
- Willson, T. M. *et al.* The structure-activity relationship between peroxisome proliferator-activated receptor γ agonism and the antihyperglycemic activity of thiazolidinediones. *J. Med. Chem.* **39**, 665–668 (1996).
- Coleman, D. L. Obese and diabetes: two mutant genes causing diabetes-obesity syndromes in mice. *Diabetologia* **14**, 141–148 (1978).
- Ingalls, A. M., Dickie, M. M. & Snell, G. D. Obese, a new mutation in the mouse. *J. Hered.* **41**, 41, 317 (1950).
- Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–432 (1994).
- Chen, H. *et al.* Evidence that the diabetes gene encodes the leptin receptor: identification of a mutation in the leptin receptor gene in db/db mice. *Cell* **84**, 491–495 (1996).
- Fujiwara, T., Yoshioka, S., Yoshioka, T., Ushiyama, I. & Horikoshi, H. Characterization of new oral antidiabetic agent CS-045. *Diabetes* **37**, 1549–1558 (1988).
- Sporst, M. B., Dunlop, N. M., Newton, D. L. & Smith, J. M. Prevention of chemical carcinogenesis by vitamin A and its synthetic analogs (retinoids). *Fed. Proc.* **35**, 1332–1338 (1976).
- Suter, S. L., Nolan, J. J., Wallace, P., Gumbiner, B. & Olefsky, J. M. Metabolic effects of new oral hypoglycemic agent CS-045 in NIDDM subjects. *Diabetes Care* **15**, 193–203 (1992).

- Mukherjee, R., Jow, L., Croson, G. E. & Paterniti, J. R. Identification, characterization, tissue distribution of human PPAR γ 2 versus PPAR γ 1 and activation with RXR agonists and antagonists. *J. Biol. Chem.* **272**, 8071–8076 (1997).
- Umesono, K., Giguere, V., Glass, C. K., Rosenfeld, M. G. & Evans, R. M. Retinoic acid and thyroid hormone induce gene expression through a common responsive element. *Nature* **336**, 262–265 (1988).
- Berger, T. S., Parandoosh, Z., Perry, B. W. & Stein, R. B. Interaction of glucocorticoid analogues with the human glucocorticoid receptor. *J. Steroid Biochem. Mol. Biol.* **41**, 733–738 (1992).
- Trinder, D. Determination of glucose in blood using glucose oxidase with an alternative oxygen acceptor. *Ann. Clin. Biochem.* **6**, 24–25 (1969).
- Mondon, C. E., Dolkas, C. B. & Oyama, J. Enhanced skeletal muscle insulin sensitivity in year-old rats adapted to hypergravity. *Amer. J. Physiol.* **240**, E482–E488 (1981).

Acknowledgements. We thank D. W. Robertson for his support and encouragement throughout these studies, E. Ulm, R. Chandraratna, J. Olefsky and P. Tontonoz for helpful discussions, L. Zhang, G. Croston, S. Lazarchik and the staff at METABOLEX Inc. for their experimental support and C. Foster for her patience and expertise in helping to prepare this manuscript.

Correspondence and requests for materials should be addressed to R.A.H. (e-mail: rheyman@ligand.com).

Biologically erodable microspheres as potential oral drug delivery systems

Edith Mathiowitz*, Jules S. Jacob*, Yong S. Jong*, Gerardo P. Carino*, Donald E. Chickering†, Pravin Chaturvedi†, Camilla A. Santos*, Kavita Vijayaraghavan*, Sean Montgomery*, Michael Bassett* & Craig Morrell*

* Department of Molecular Pharmacology, Physiology and Biotechnology, Division of Biology and Medicine, Brown University, Providence, Rhode Island 02912, USA

Biologically adhesive delivery systems offer important advantages^{1–5} over conventional drug delivery systems⁶. Here we show that engineered polymer microspheres made of biologically erodable polymers, which display strong adhesive interactions with gastrointestinal mucus and cellular linings, can traverse both the mucosal absorptive epithelium and the follicle-associated epithelium covering the lymphoid tissue of Peyer's patches. The polymers maintain contact with intestinal epithelium for extended periods of time and actually penetrate it, through and between cells. Thus, once loaded with compounds of pharmacological interest, the microspheres could be developed as delivery systems to transfer biologically active molecules to the circulation. We show that these microspheres increase the absorption of three model substances of widely different molecular size: dicumarol, insulin and plasmid DNA.

Although hydrophilic polymers and hydrogels containing carboxyl groups have traditionally been considered to be the best biological adhesives^{1–5}, fast-degrading hydrophobic polymer microspheres containing surface carboxylic acid groups display the same properties^{7,8}. Indeed, in a comparative study, polyanhydride copolymers of fumaric and sebacic acid, poly (FA:SA), demonstrated the highest biologically adhesive properties⁹. *In vitro* up to 65% of a batch of microspheres made of these copolymers adhered spontaneously to everted intestinal sacs¹⁰. *In vivo* radiographic studies on these microspheres showed that they remained in the stomach longer than microspheres made of other polymers (36 h against 18 h)^{11,12}. These results suggest that strong biologically adhesive interactions delay the passage of the microspheres through the gastrointestinal system.

It has been demonstrated that small microspheres of polystyrene, poly-lactide-co-glycolide and polycyanoacrylate could be taken up by lymphoid tissue associated with Peyer's patches in the gastrointestinal tract^{13–22}. To explore the potential of uptake of the new formulations, we used optical microscopy to follow orally administered poly(FA:SA) 20:80 microspheres (diameter 0.1–

† Present addresses: Sulzer Carbomedics, Austin, TX 78752, USA (D.E.C.); Vertex Pharmaceuticals, Cambridge, MA, USA (P.C.).

10 μm) made using phase inversion nanoencapsulation (PIN) (Fig. 1). As early as 1 h after feeding of a single dose, microspheres were observed to traverse both the mucosal epithelium through and between individual cells and the follicle-associated epithelium covering the lymphatic elements of Peyer's patches. They were also observed in goblet cells and crossing through follicle associated epithelium and into Peyer's patches. A large increase in uptake was observed in absorptive cells and Peyer's patches after 3, 6, 12 and 24 h (Fig. 2a). Both spleen and liver tissue samples, with apparently normal-looking hepatocytes, showed a large number of microspheres (Fig. 2b); similar patterns were never observed in sections of tissue from untreated control animals.

Transmission electron microscopy (TEM) using electron-opaque, 5-nm colloidal gold tracer that had been PIN-encapsulated with the new copolymer (Fig. 3a–d) revealed that the range of particle sizes observed in the cytoplasm of cells was 40–120 nm, well below the resolution of normal light optics (Fig. 3a–d). Colloidal gold-labelled nanospheres were visualized in the cytoplasm, inside Golgi apparatus, and in secretory vesicles (Fig. 3a, d, e), predominantly in the supranuclear (apical) portion of the absorptive cells (Fig. 3b). The mechanism of entry was not clear, as relatively few pinocytotic vesicles were observed to account for the uptake. Microspheres were often found between microvilli, in vesicles near the lateral borders of the cell (Fig. 3d), in the intercellular spaces (Fig. 3a, d), in the basal aspects of the cells, in lymphatics, and in goblet cells (data not shown). Unequivocal identification of the particles was difficult because the microspheres had different appearances depending upon their location in the cell or intercellular spaces, perhaps because of degradation or processing of the polymer. However, the particles were identified on the basis of the colloidal gold tag. Microspheres were never observed in control tissues from untreated animals (Fig. 3f). These findings suggest that translocation of polymer particles through the transcellular route, as well as through paracellular transport, had occurred. A similar study using colloidal gold-labelled polystyrene showed high uptake, whereas microspheres made of polylactic acid, a polyester shown to be much less biologically adhesive than poly(FA:SA) 20:80, showed only minimal uptake.

The positive uptake results led us to determine whether biologically adhesive microspheres could be used to carry and increase the absorption of three model drugs: dicumarol, insulin and plasmid DNA. Dicumarol, a low-molecular-weight anticoagulant drug with erratic gastrointestinal uptake and low solubility^{23,24}, was encapsulated using the PIN method to produce particles with diameters less than 5 μm (80% of the spheres were <1 μm , and the remaining 20% were between 1 and 4 μm ; see Fig. 1). Uptake of encapsulated dicumarol was compared with that of a stock suspension of ordinary dicumarol and spray-dried unencapsulated drug (to reduce the drug particle size to a range similar to that of the PIN spheres). The plasma area under the curve (AUC) for the encapsulated formulation was found to be $363.3 \mu\text{g h}^{-1} \text{ml}^{-1}$, compared with $232.1 \mu\text{g h}^{-1} \text{ml}^{-1}$ for spray-dried drug ($P < 0.034$) and $171.5 \mu\text{g h}^{-1} \text{ml}^{-1}$ for stock dicumarol ($P < 0.012$). The relative calculated bioavailability for the three formulations was 19% for the poly(FA:SA)-encapsulated microspheres, 12.2% for the spray-dried dicumarol, and 8.9% for the stock dicumarol. Thus the poly(FA:SA)-encapsulated drug formulations demonstrated a 57% increase in the AUC plasmas concentration–time curve over the spray-dried, unencapsulated drug, and a 112% increase over the unmiconized stock suspension. Because the oral bioavailability of dicumarol in the microspheres was estimated to be 19%, and the microspheres were loaded with 40% dicumarol, up to 47% of the microparticles were estimated to be absorbed in the rats in the study. Furthermore, dicumarol was found to remain in plasma for significantly longer periods than with either the spray-dried or stock suspensions¹¹, suggesting an improvement in the oral bioavailability of the drug.

Insulin was encapsulated in a blend of poly(fumaric anhydride) and poly(lactide-co-glycolide) 50:50 (poly FA:PLGA) (M_n 0.3K and 35K, respectively) using the PIN fabrication method, leading to a mean particle size of 96.7 nm with 90% of the molecules being less than 201 nm. To evaluate the gastrointestinal uptake of the insulin formulation, three groups of fasted rats were injected subcutaneously with an initial glucose load²⁵ and then fed: (1) a suspension of poly(FA:PLGA) microspheres containing 20 units zinc–insulin with micronized FeO included as an electron-dense TEM tracer in

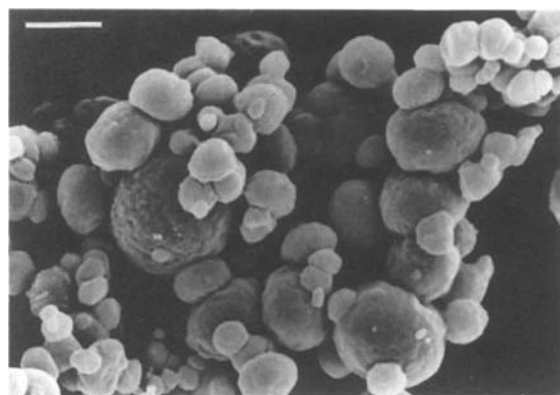


Figure 1 SEM of poly(FA:SA) 20:80 PIN microspheres prepared by dispersing 4% (w/v) polymer solution in methylene chloride into a 100-fold excess of petroleum ether (non-solvent). No agitation or stirring was used during the fabrication process. Under these conditions, microparticles in of 0.3–2.0 μm are produced with a mean diameter of 0.7 μm . Particles of different sizes can be fabricated by changing the initial polymer solution, solvent to non-solvent ratio, and selection of non-solvent. Scale bar, 5 μm .

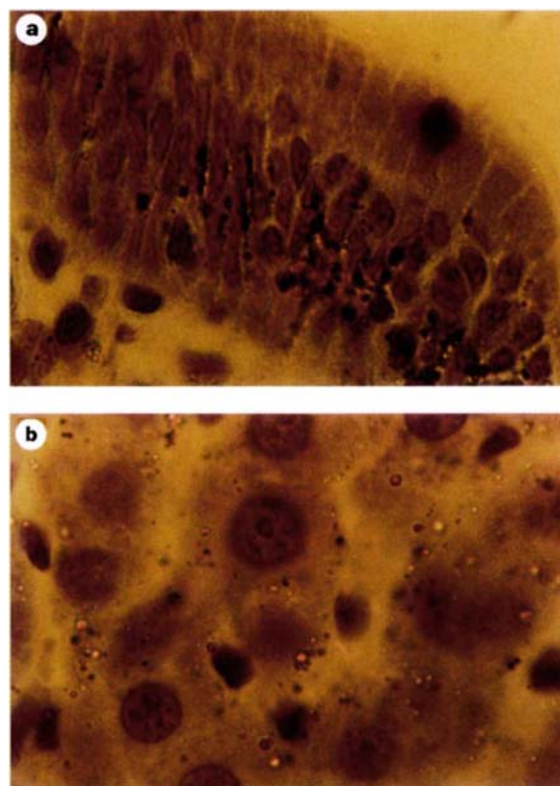


Figure 2 Histological evidence of uptake of poly(FA:SA) nanospheres across gastrointestinal epithelium. **a**, An ileal segment 24 h after feeding showing many spheres still residing in spaces between epithelial cells. **b**, A liver segment 6 h after feeding showing many spheres with normal hepatocytes. All magnifications, $\times 1,000$.

saline ($n = 6$); (2) 24 units of soluble insulin in saline ($n = 8$); or (3) saline without insulin as controls ($n = 7$).

The control rats showed the expected response to the glucose load (Fig. 4). In saline-fed animals, blood glucose levels rose by 46 mg dl^{-1} after 3 h, then slowly returned to baseline. Rats fed soluble insulin reached a maximum blood glucose level of 36 mg dl^{-1} above baseline at 1.5 h post-feeding. At all time points, the blood glucose levels of rats receiving the insulin solution were not significantly different from those of control animals receiving

saline only. In contrast, the blood glucose levels of animals fed the PIN-encapsulated insulin formulation never increased above the fasting levels, and were statistically lower than both control groups at all four time points (see Fig. 4). Clearly, animals fed the poly(FA:PLGA)-encapsulated insulin preparation were better able to regulate the glucose load than the controls, suggesting that the insulin crossed the intestinal barrier and was released from the microspheres in a biologically active form.

Formulations of blends of poly(LA) and poly(FA), up to ratios of

Figure 3 Transmission electron microscopy (TEM).

a, Three absorptive cells near the tip of a jejunal villus, from a fasted rat fed 20 mg of poly(FA:SA) 20:80 PIN microspheres containing colloidal gold, killed after 6 h showing, lumen (lum). Even at this magnification, particles were observed in the cytoplasm and intercellular spaces. Region A, the apical portion of the cell below the terminal web; region B, the supranuclear region. Scale bar, $1 \mu\text{m}$. **b**, A higher-magnification image of region A, showing numerous small particles containing colloidal gold (arrows), immediately below the microvilli (Mv). Scale bar, $1 \mu\text{m}$. **c**, Region B with nucleus (Nu). Particles were visible throughout the cytoplasm, in dilated Golgi apparatus (asterisk) and secretory vesicles near the lateral edges of the cell, and in intercellular spaces (IC). Scale bar, $1 \mu\text{m}$. **d**, Region C from **c**. The lysosome (L) can also be seen in **a**, **c** and **d**. Scale bar, $1 \mu\text{m}$. **e**, Six particles from region D of **d**. Electron-dense structures within the particles, with approximate dimensions of colloidal gold, are probably clusters of individual colloidal gold particles. Scale bar, 100 nm . **f**, Four absorptive cells from an untreated, control rat, showing rough endoplasmic reticulum (rer) and empty Golgi apparatus unlike cells from the experimental animals. Structures resembling microspheres were not observed. Scale bar, $1 \mu\text{m}$. All sections were either unstained or very lightly stained with uranyl acetate to minimize artefacts and allow observation of colloidal gold.

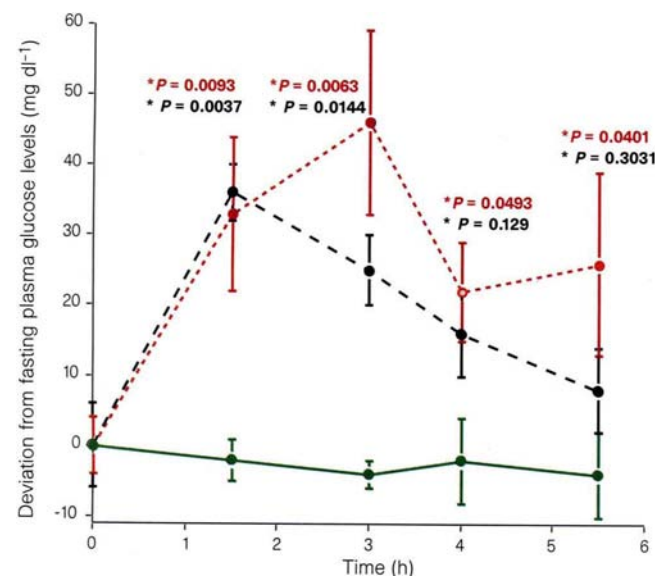
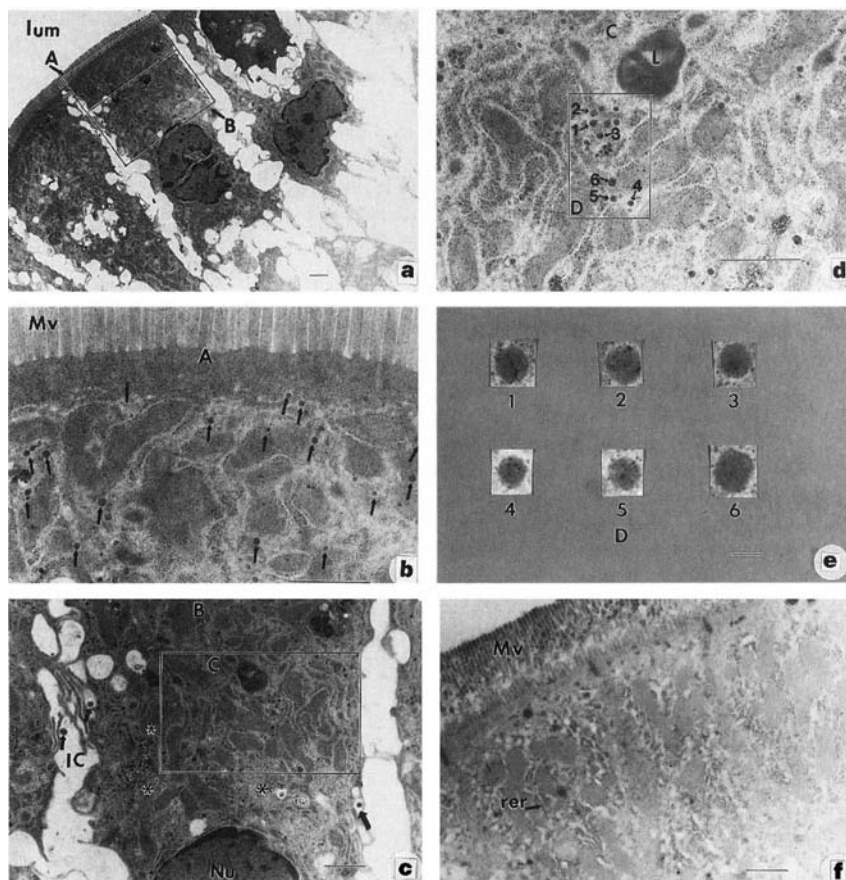


Figure 4 Reaction to glucose load in rats fed 1.5% (w/w) insulin-loaded poly(FA:PLGA) PIN microspheres (green, $N = 6$), 20 IU soluble insulin in saline (black, $N = 8$), and saline buffer (red, $N = 7$). All data are mean $\pm$ s.e.

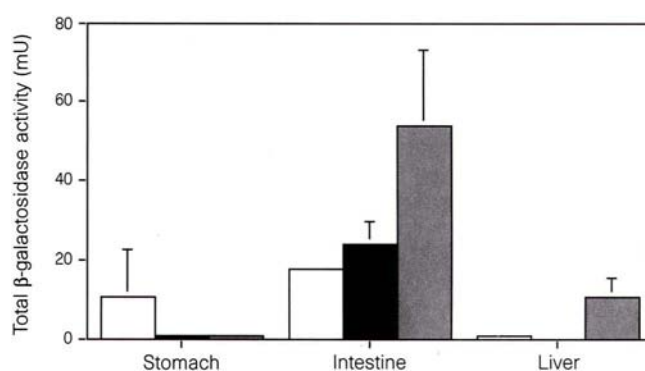


Figure 5 Quantification of β -galactosidase activity in tissue extracts. Results are expressed as mU of β -galactosidase activity for the entire tissue homogenate. All data are mean $\pm$ s.e. Background activity (white, $N = 4$); plain pCMV/ β gal (black, $N = 3$); poly(FA:SA)-encapsulated pCMV/ β gal (grey, $N = 3$). Background activity was determined by similar digests of tissues from unfed rats.

1:3, loaded with the same amount of insulin (with similar drug release kinetics, biological and particle size) did not show any effect in the glucose tolerance test (results not shown). This observation correlated well with the low level of particle uptake using poly(LA). The differences observed between the two polymer formulations may result from either the less biological adhesive properties or slower degradation rate of the blends of the polymers.

Perhaps the most surprising results were found when we applied the system to the delivery of plasmid DNA for *in vivo* gene transfer. Because 'naked' plasmid DNA transfection efficiencies are generally low as a result of inefficient uptake as well as rapid degradation, encapsulation of plasmid DNA might serve to protect DNA and promote cellular uptake after oral administration. A biologically adhesive polymer delivery system composed of poly(FA:SA) in a 20:80 molar ratio was used to encapsulate pCMV/ β -gal (0.1%), a commercially available reporter plasmid with bacterial β -galactosidase gene activity. The migration pattern of plasmid DNA extracted from the microspheres was identical to stock unencapsulated DNA, indicating that the fabrication process did not alter its structural integrity (data not shown).

Five days after a single oral dose of either plasmid-loaded PIN microspheres or unencapsulated pCMV/ β -gal, β -galactosidase activity was quantified in the stomach, small intestine and liver (Fig. 5). Animals fed encapsulated pCMV/ β -gal showed greater levels of β -galactosidase activity in both the small intestine and the liver than either those receiving unencapsulated material or unfed animals. The reporter gene activity measured in intestinal tissue of animals that received the encapsulated pCMV/ β -gal was highest (>54 mU versus 24 mU for the unencapsulated plasmid and 18 mU for the background levels in untreated control animals; however, there was no significant difference between the three groups). These same animals averaged 11 mU of activity in the liver versus <1 mU for plain CMV-fed ($P < 0.045$) or untreated control animals ($P < 0.025$). Reporter gene expression in stomach homogenates was generally low (<11 mU) in all groups.

In another cohort of animals, visual localization of transfected cells after oral administration was performed using X-gal histochemical staining techniques on whole tissue and frozen sections²⁶. Because epithelial cells on villi have endogenous lactase activity,

which can mask the bacterial β -galactosidase²⁷, we focused on Peyer's patches, which do not contain background activities.

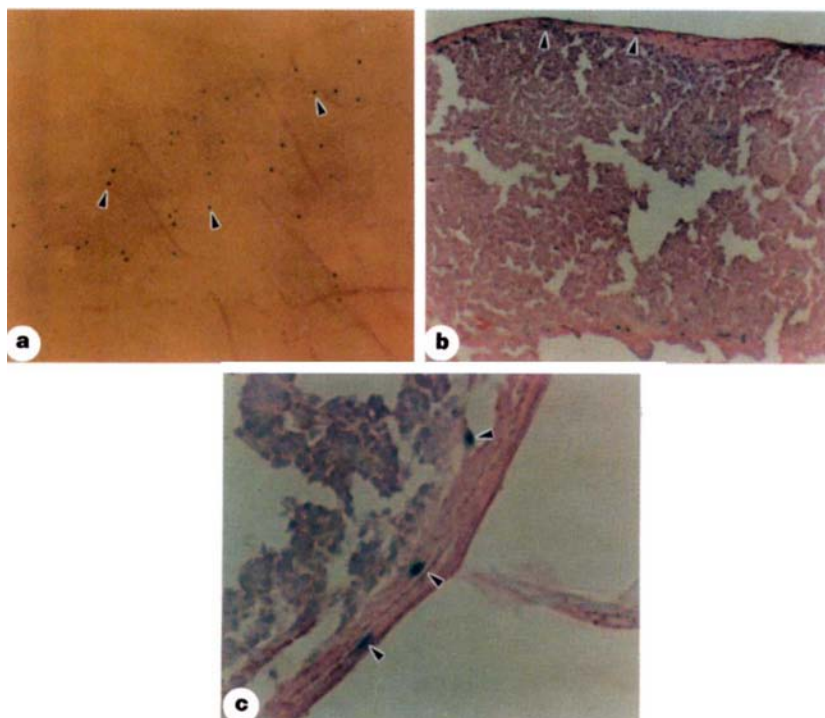
The serosal surface of small intestine from encapsulated pCMV/ β -gal-fed rats showed intensely stained cells in areas containing Peyer's patches (Fig. 6a). Analysis of frozen sections of Peyer's patches showed that, although there were a few β -galactosidase positive cells within the central lymphoid tissue mass (Fig. 6b), most of the transfected cells were located in the muscularis mucosae and adventitia below the Peyer's patches (Fig. 6c). This staining distribution was consistent with previous studies showing retention of microspheres in the Peyer's patches. Neither the animals receiving unencapsulated pCMV/ β -gal nor unfed normal rats showed any false-positive β -galactosidase staining in the region of the Peyer's patches. Histological examination showed no evidence of mucosal damage or inflammation. This study confirmed that plasmid DNA can be delivered by the oral route using PIN formulations, with the encapsulated DNA incorporated into cells lining the small intestine and hepatocytes, and expressing functional gene products at levels detectable using common histological and luminometric techniques.

In conclusion, using the PIN system in combination with biologically erodable, biologically adhesive polymers, the enhanced bioavailability of dicumarol might be explained by the delay in the dissolution of the drug caused by the encapsulation process¹¹, the increased absorption rate resulting from more intimate contact with the intestinal epithelium, or an increased residence time at the local absorption site (all of which can be attributed to 'biological adhesion'). However, the increased biological activity of insulin and plasmid DNA might also be due to the uptake of the microspheres into the circulation by the cells lining the gastrointestinal tract. The results with insulin, a 6K protein susceptible to proteolytic degradation by gastrointestinal enzymes and having limited mucosal uptake, support this hypothesis. The same uptake pathways can be used for a variety of therapeutic agents with poor oral absorption, including nucleic acids and plasmids. □

Methods

Polymer synthesis. Poly(FA:S) 20:80 (M_n 0.5K to 2K) was synthesized by melt condensation polymerization of fumaric acid and sebacic acid²⁸.

Figure 6 X-gal staining for the histological localization of transfected cells in animals fed with poly(FA:SA)-encapsulated pCMV/ β -gal. **a**, Whole-tissue X-gal staining of the serosal surface of a Peyer's patch area in the small intestines. Many positively stained, individual, blue cells are apparent. **b**, X-gal staining of a Peyer's patch area (20- μ m cryostat section) in the small intestine, showing positive staining below the lymphoid cells of the Peyer's patch, as well as positive cells in the core that were never observed in controls. **c**, A similar tissue section at a higher magnification, showing three transfected cells, apparently in the muscularis mucosae.



Microencapsulation by phase inversion. Drug was added to a dilute polymer solution (1–4% w/v in methylene chloride), which was then poured rapidly into an unstirred bath of non-solvent (petroleum ether) at a solvent to non-solvent ratio of 1:100, causing nano and microspheres (0.1–5.0 μm in diameter) to form spontaneously.

Bioavailability studies. The jugular veins of rats were catheterized, and serum dicumaryl levels were taken after oral feeding of different drug formulations¹¹.

Histological studies. Rats fasted overnight were fed a single dose (3 mg) of PIN poly(FA:SA) 20:80 in 1 ml of saline, and killed after 1, 3, 6, 12 or 24 h. Segments of intestine, liver and spleen were excised, fixed in 2% paraformaldehyde, 0.2% glutaraldehyde in 0.1 M PBS, pH 7.4, dehydrated in graded ethanols, embedded in glycomethacrylate, and cut into 5- μm sections. Light-microscopy samples were stained with cresyl violet. X-gal histochemistry was performed on cryostat section and whole tissue²⁹. For TEM, the tissue and microspheres were additionally fixed with 1% osmium tetroxide in 0.1 M PBS, pH 7.4, before dehydration, infiltrated with Spurr's epoxy medium, embedded and cured at 60 °C for 20 h. Control microspheres were processed in a manner identical to the tissue. These sections (80–90 nm) were stained very lightly with uranyl acetate for 1 min and observed at 80 kV, or were observed unstained.

Glucose tolerance test. Baseline blood glucose levels were assayed in rats fasted overnight which then received a subcutaneous glucose load of 800 mg per kg. Experimental rats ($n = 6$) were fed 20 units zinc–insulin encapsulated in PIN nanospheres. Two cohorts of control rats were used: rats fed 1 ml of saline ($n = 7$), rats fed 20 units of soluble insulin in 1 ml of saline ($n = 8$). At 1.5, 3, 4 and 5.5 h post-glucose load, blood samples were collected for determination of glucose level.

Insulin release. Studies were performed in 0.1 M phosphate buffer, pH 7.4, at 37 °C and analysed for protein using the BCA assay. Poly(FA):PLGA 50:50 nanospheres containing 1.5% insulin (w/w) were incubated in buffer and the incubation fluids were assayed; 40% of the total loading was released after 30 min; 58% after 1 h; 65% after 5 h; 82% after 40 h; and 93% after 73 h. SDS–PAGE of the incubation fluids demonstrated a single protein band migrating at an apparent M_r of 6K. In contrast, nanospheres fabricated with only poly(FA:SA) 20:80 released 90% of the insulin load within 30 min and were not suitable.

Insulin biological activity. Insulin-encapsulated PIN formulations were tested for biological activity by intraperitoneal injection of 50 mg of microspheres in saline into fasted rats. Typically, fasting blood glucose levels declined by 45–50% after 1.5 h and 90–95% after 4 h. Only biologically active formulations were used for oral feeding.

Quantification of transgene activity. Unencapsulated plasmid (500 μg) (control) or 50 μg plasmid encapsulated in poly(FA:SA) 20:80 (M_n 2K) were administered by stomach tube as a single dose to fasted 400-g rats and compared with unfed normal rats. The encapsulated plasmid dosing was given at 10% of the control plasmid dose to test the efficacy of the biologically adhesive delivery system and to demonstrate the protective benefits of encapsulation. After 5 days the rats were killed, and the stomach, small intestine and liver were excised and tested for β -galactosidase expression using a chemiluminescent substrate, Galacto-Light (Tropix, Bedford, MA)³⁰. Background activity was determined by analysing similar digests of tissues from unfed rats.

Received 12 July 1996; accepted 31 January 1997.

1. Lenaerts, V., Couvreur, P., Grislain, L. & Maincent, P. in *Bioadhesive Drug Delivery Systems* (eds Lenaerts, V. & Gurny, R.) 93–104 (CRC Press, Boca Raton, FL, 1990).
2. Peppas, N. A. *Hydrogels in Medicine and Pharmacy* (CRC Press, Boca Raton, FL, 1987).
3. Ponchel, G. & Duchêne, D. in *High Performance Biomaterials* (ed. Szycher, M.) 231–242 (Technomic, Lancaster, PA, 1991).
4. Davis, S. S. & Illum, L. *Biomaterials* **9**, 111–115 (1988).
5. Frišbie, C. D., Lawrence, F., Rozsnyai, A. N., Wrighton, M. S. & Charles, M. L. *Science* **265**, 2071–2073 (1994).
6. Langer, R. & Chasin, M. *Biodegradable Polymers as Drug Delivery Systems* (Marcel Dekker, New York, 1990).
7. Chickering, D. E., Jacob, J. S. & Mathiowitz, E. *React. Polym.* **25**, 189–206 (1995).
8. Mathiowitz, E., Jacob, J. S., Chickering, D. & Pekarek, K. *Macromolecules* **26**, 6756–6765 (1993).
9. Chickering, D. E. & Mathiowitz, E. *J. Control. Release* **34**, 251–261 (1995).
10. Jacob, J., Santos, C., Carino, G., Chickering, D. & Mathiowitz, E. *Proc. 22nd Int. Symp. Controlled Release of Bioactive Materials*, 312–313 (Controlled Release Society, Inc., Deerfield, IL, 1995).
11. Chickering, D. E. III, Jacob, J. S. & Mathiowitz, E. *Biotechnol. Bioeng.* **52**, 96–101 (1996).
12. Carino, G., Santos, C., Chen, C. J. & Mathiowitz, E. *Proc. 22nd Int. Symp. Controlled Release of Bioactive Materials*, 314–315 (Controlled Release Society, Inc., Deerfield, IL, 1995).
13. Sanders, E. & Ashworth, C. T. *Exp. Cell Res.* **22**, 137–145 (1961).
14. Ermak, T. H., Dougherty, E. P., Bhagat, H. R., Kabok, Z. & Pappo, J. *Cell Tissue Res.* **279**, 433–436 (1995).

15. Florence, A. & Jani, P. U. *Drug Safety* **10**, 233–266 (1994).
16. Florence, A. T., Hillery, A. M., Hussain, N. & Jani, P. U. *J. Control. Release* **36**, 39–46 (1995).
17. Lefevre, M. E., Olivo, R., Vanderhoff, J. W. & Joel, D. D. *Proc. Soc. Exp. Biol. Med.* **159**, 298–302 (1978).
18. Lefevre, M. E., Joel, D. D. & Schidlovsky, G. *Proc. Soc. Exp. Biol. Med.* **179**, 522–528 (1985).
19. Jani, P., Halbert, G. W., Langridge, J. & Florence, A. T. *J. Pharm. Pharmacol.* **41**, 809–812 (1989).
20. Jani, P., Halbert, G. W., Langridge, J. & Florence, A. T. *J. Pharm. Pharmacol.* **42**, 821–826 (1990).
21. Hillery, A. M., Jani, P. U. & Florence, A. T. *J. Drug Target.* **2**, 151–156 (1994).
22. Bockman, D. E. & Cooper, M. D. *Am. J. Anat.* **136**, 455–478 (1973).
23. Akers, M. J., Lach, J. L. & Fischer, L. J. *J. Pharma. Sci.* **62**, 391–395 (1973).
24. Goodman, L. S. & Gilman, A. *The Pharmacological Basis of Therapeutics* (Macmillan, New York, 1975).
25. Tutwiler, G. F., Kirsch, T. & Bridi, G. *Diabetes* **27**, 856–867 (1978).
26. Wolff, J. A. *et al. Biotechniques* **11**, 474–485 (1991).
27. Lau, C. *et al. Hum. Gene Ther.* **6**, 1145–1151 (1995).
28. Domb, A., Mathiowitz, E., Ron, E., Giannos, S. & Langer, R. J. *Polym. Sci.* **29**, 571–579 (1991).
29. Wolff, J. A. *et al. Science* **247**, 1465–1468 (1990).
30. Lee, S. W., Trapnell, B. C., Rade, J. J., Virmani, R. & Dichek, D. A. *Circ. Res.* **73**, 797–807 (1993).

Acknowledgements. We thank P. Galletti for analysis of the manuscript. This work was partially supported by an NIH grant.

Correspondence and requests for materials should be addressed to E.M. (e-mail: Edith_Mathiowitz@Brown.edu).

An atypical topoisomerase II from archaea with implications for meiotic recombination

Agnès Bergerat*, Bernard de Massy†, Danielle Gadelle*, Paul-Christophe Varoutas†, Alain Nicolas† & Patrick Forterre*

* Institut de Génétique et Microbiologie, Université Paris Sud, CNRS URA 1354, GDR 1006, 91405 Orsay Cedex, France

† Institut Curie, Section de Recherche, CNRS UMR144, 26 rue d'Ulm, 75248 Paris Cedex 05, France

Type II topoisomerases help regulate DNA topology during transcription, replication and recombination by catalysing DNA strand transfer through transient double-stranded breaks¹. All type II topoisomerases described so far are members of a single protein family². We have cloned and sequenced the genes encoding the A and B subunits of topoisomerase II from the archaeon *Sulfolobus shibatae*. This enzyme is the first of a new family. It has no similarity with other type II topoisomerases, except for three motifs in the B subunit probably involved in ATP binding and hydrolysis. We also found these motifs in proteins of the Hsp90³ and MutL⁴ families. The A subunit has similarities with four proteins of unknown function. One of them, the *Saccharomyces cerevisiae* Spo11⁵ protein, is required for the initiation of meiotic recombination. Mutagenesis, performed on *SPO11*, of the single tyrosine conserved between the five homologues shows that this amino acid is essential for Spo11 activity. By analogy with the mechanism of action of known type II topoisomerases, we suggest that Spo11 catalyses the formation of double-strand breaks that initiate meiotic recombination in *S. cerevisiae*.

Topoisomerases change the degree of superhelicity, catenation and knotting of topologically closed DNA molecules. They are classified into type I and II according to the nature of this transient break, single- or double-stranded, respectively¹. Type I topoisomerases, which are mostly ATP-independent, are divided into two families, A and B, which have no sequence similarities between them². In contrast, all type II topoisomerases are ATP-dependent and have many similarities at the amino-acid sequence level². Bacteria possess two type II topoisomerases, gyrase and topoisomerase IV (the fourth topoisomerase described in *Escherichia coli*). These enzymes are heterotetramers composed of two subunits Gyr A and Gyr B, and Par E and Par C, respectively. Eukaryotes possess only one or two isoforms of a single topoisomerase II whose amino- and carboxy-terminal domains are, respectively, homologues of Gyr B (Par E) and Gyr A (Par C). Gyr B and homologues

tion. In *S. cerevisiae*, meiotic recombination is initiated by the formation of double-strand breaks (DSBs). These DSBs are processed by resection of the 5' ends, and repaired by recombination with the homologous chromosome in wild-type diploids¹⁷. In diploids homozygous for the *rad50S* mutation, no resection is detected, and DSBs accumulate¹⁸. Recently, a protein that is covalently bound to the 5' ends of the DSBs in *rad50S* strains has been found^{19–21}. As this is reminiscent of the mechanism of a topoisomerase II, the authors suggested that a topoII-like activity is responsible for the formation of meiotic DSBs in *S. cerevisiae*. Mutations in several genes result in the absence of meiotic DSBs and meiotic recombination²². *SPO11* is one of these genes²³.

In all topoisomerases, DNA cleavage is initiated by the attack on the phosphodiester bond by a tyrosine residue to form a covalent DNA–protein transient intermediate¹. We noticed that a single tyrosine residue (tyrosine 135 in Spo11) is common to topo VI and all their eukaryotic homologues in our alignment (Fig. 1). This tyrosine could be critical for Spo11, and possibly mediates DNA cleavage. We have replaced by site-directed mutagenesis this tyrosine by a phenylalanine. Diploids homozygous for this *spo11YF135* allele were obtained and tested for various phenotypes in comparison with wild-type strains and strains carrying a homozygous disruption of *SPO11* (*spo11::URA3*). Meiotic DSBs appear as transient DNA fragments in the wild-type strain (*SPO11/SPO11*), but are not detected in both the *spo11::URA3/spo11::URA3* and *spo11YF135/spo11YF135* strains (Fig. 3). In a similar manner to the *spo11::URA3/spo11::URA3* strain, the *spo11YF135/spo11YF135* strain showed no induction of meiotic recombination in a return to growth experiment, reduced spore formation and low spore viability (data not shown).

These results show that the tyrosine 135 of Spo11 is essential for its activity and support our sequence alignment, identifying Spo11 as a homologue of the *S. shibatae* topo VI A subunit. We thus propose that Spo11 generates the DSBs that initiate meiotic recombination by a cleavage reaction similar to the one catalysed by archaeal topo VI. The sequence similarities between Rec12 of *S. pombe* and Spo11 of *S. cerevisiae* (Fig. 1) and the similarity of phenotypes of mutants in *REC12* and *SPO11* genes^{5,10}, suggests that Rec12 is also involved in the initiation of meiotic recombination by forming DSBs in *S. pombe*.

The homology between Spo11 and the A subunit of archaeal topo VI also raises the question of the existence of a new topoisomerase II in eukaryotes: whether Spo11 interacts with another subunit to

generate a topoisomerase activity. Note that *SPO11* is also required for homologous chromosome pairing during meiosis²⁴. A topoisomerase activity could, for example, play a direct role in chromosome pairing by the stabilization of homologous interactions. There is no close homologue of *S. shibatae* topo VI B subunit in the complete yeast genome²⁵, however, it might be significant that Hsp82, a yeast member of the Hsp90 family, is induced during meiosis²⁶.

The discovery of a new family of type II topoisomerases in the Archaea was unexpected and opens a new field of research in DNA topoisomerases. The definition of a new ATP module should have important implications in the study of Hsp90 and MutL. Finally, the detection of homology between *S. shibatae* topoisomerase II and proteins involved in meiotic recombination emphasizes that studying archaea can help to clarify critical steps in eukaryotic molecular biology. □

Note added in proof: Our results predict that Spo11 is covalently bound to meiotic DSBs in *Rad50S*. This association has recently been demonstrated²⁷.

Methods

Cloning and sequencing. A fragment of 413 bp was amplified by PCR using degenerated primers designed from the two peptidic sequences KAWQEEINT and AVFQQLHRAG (single-letter amino-acid code) of the *S. shibatae* topo VI A subunit. This probe was used to screen *S. shibatae* genomic DNA, after digestion with *Hind*III, by Southern blot hybridization. A fragment of about 3 kb was cloned in pUC18 and sequenced. This fragment contained the complete gene for the A subunit. An open reading frame (ORF) located upstream of this gene appeared to encode the C-terminal part of the B subunit, because it contains one of the peptidic sequences (ELVSKYQNEISEGLF) obtained from this subunit. This partial ORF was used as a probe to clone the entire region containing the B subunit gene.

Mutagenesis. Mutagenesis of the *SPO11* gene was done with a site-directed mutagenesis kit (QuickChange, Stratagene) on a plasmid carrying the *Hind*III–*Hind*III fragment of *SPO11*. The TAC codon for tyrosine 135 was substituted by a TTC codon (phenylalanine). After mutagenesis, the entire *SPO11* ORF was sequenced on both strands and was found to have the correct substitution and no other mutation. A one base deletion in a row of 10 Ts in the non-coding region 240 bp upstream of the *SPO11* ATG was observed.

Strain constructions, genotypes and analysis. The *spo11::URA3* disruption is a replacement of most of the *SPO11* ORF by the *URA3* gene, introduced in *SPO11* haploid strains by gene replacement. The *spo11YF135* allele was introduced in haploid strains carrying the *spo11::URA3* disruption. All strains were verified by PCR and Southern blot analysis. ORD887 (*a/α*, *poly1arg4RV/poly1arg4Bg*, *ADE2/ade2*, *cyhr/cyhr*, *HIS3/his3Δ1*, *LEU2/leu2-3,112*, *trp1-289/trp1-289*, *ura3-52/ura3-52*), ORD3021 (*a/α*, *poly1arg4RV/poly1arg4Bg*, *ADE2/ade2*, *cyhr/cyhr*, *his3Δ1/his3Δ1*, *leu2-3,112/leu2-3,112*, *trp1-289/trp1-289*, *ura3-52/ura3-52*, *spo11::URA3/spo11::URA3*), ORD4209 (*a/α*, *poly1arg4RV/poly1arg4Bg*, *ADE2/ade2*, *cyhr/cyhr*, *his3Δ1/his3Δ1*, *leu2-3,112/leu2-3,112*, *trp1-289/trp1-289*, *ura3-52/ura3-52*, *spo11YF135/spo11YF135*). Sporulation and Southern blot analysis were carried out as described previously²⁸.

Received 19 November 1996; accepted 31 January 1997.

- Wang, J. C. DNA topoisomerases. *Annu. Rev. Biochem.* **65**, 635–692 (1996).
- Caron, P. L. & Wang, J. C. Alignment of primary sequences of DNA topoisomerases. In *Advances in Pharmacology* (ed. Liu, L. F.) 271–297 (Biology Academic, New York, 1994).
- Gupta, R. S. Phylogenetic analysis of the 90 kDa heat shock family of protein sequences and an examination of the relationship among animal, plants and fungi species. *Mol. Biol. Evol.* **12**, 1063–1073 (1995).
- Papadopoulos, N. *et al.* Mutation of a mutL homolog in hereditary colon cancer. *Science* **263**, 1625–1629 (1994).
- Atcheson, C. L., DiDomenico, B., Frackman, S., Esposito, R. E. & Elder, R. T. Isolation, DNA sequence, and regulation of a meiosis-specific eukaryotic recombination gene. *Proc. Natl Acad. Sci. USA* **84**, 8035–8039 (1987).
- Woese, C. R. & Fox, G. E. Phylogenetic structure of prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090 (1977).
- Homes, M. L., Nuttall, S. D. & Dyll-Smith, M. L. Construction and use of halobacterial shuttle vectors and further studies on *Halobacterium* DNA gyrase. *J. Bacteriol.* **173**, 3807–3813 (1991).
- Bergerat, A., Gadelle, D. & Forterre, P. Purification of DNA topoisomerase II from the hyperthermophilic Archaeon *Sulfolobus shibatae*. *J. Biol. Chem.* **269**, 27663–27669 (1994).
- Wilson, R. *et al.* 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368**, 32–38 (1994).

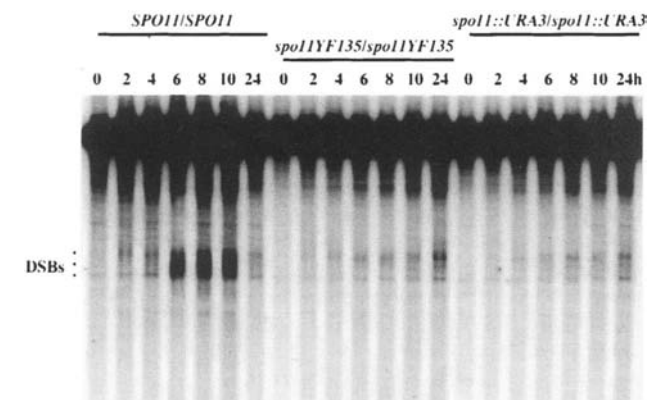


Figure 3 Absence of meiotic DSBs in strains carrying the *spo11YF135* allele. Three strains were analysed in parallel: a wild-type (ORD887, *SPO11/SPO11*), the *spo11YF135* mutant (ORD4209, *spo11YF135/spo11YF135*) and a mutant disrupted in the *SPO11* gene (ORD3021, *spo11::URA3/spo11::URA3*). Chromosomal DNA was extracted at various times after transfer to sporulation medium, digested with *Hind*III, and analysed by Southern hybridization to detect DSBs in the promoter region of the *CYS3* gene as described²⁷. The meiotic DSBs appear as a smear at 6, 8 and 10 h in the wild-type strain (*SPO11/SPO11*).

10. Lin, Y. & Smith, G. R. Transient, meiosis-induced expression of *rec6* and *rec12* genes of *Schizosaccharomyces pombe*. *Genetics* **136**, 769–779 (1994).
11. Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*. *Science* **273**, 1058–1072 (1995).
12. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. *Proc. Natl Acad. Sci. USA* **87**, 4576–4579 (1990).
13. Wigley, D. B., Davies, G. J., Dodson, E. J., Maxwell, A. & Dodson, G. Crystal structure of an N-term fragment of the DNA gyrase protein. *Nature* **351**, 624–629 (1991).
14. Jackson, A. P. & Maxwell, A. Identifying the catalytic residue of the ATPase reaction of DNA gyrase. *Proc. Natl Acad. Sci. USA* **90**, 11232–11236 (1993).
15. Louvion, F. L., Warth, R. & Picard, D. Two eukaryotic-specific regions of Hsp82 are dispensable for its viability. *Proc. Natl Acad. Sci. USA* **93**, 13937–13942 (1996).
16. Aronshtam, A. & Marinus, M. G. Dominant negative mutator mutations in the *mutL* gene of *E. coli*. *Nucleic Acids Res.* **24**, 2498–2504 (1996).
17. Sun, H., Treco, D. & Szostak, J. W. Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the *ARG4* recombination initiation site. *Cell* **64**, 1155–1161 (1991).
18. Alani, E., Padmore, R. & Kleckner, N. Analysis of wild-type and *rad50* mutants of yeast suggests an intimate relationship between meiotic chromosome synapsis and recombination. *Cell* **61**, 419–436 (1990).
19. de Massy, B., Rocco, V. & Nicolas, A. The nucleotide mapping of DNA double-strand breaks at the *CYS3* initiation site of meiotic recombination in *Saccharomyces cerevisiae*. *EMBO J.* **14**, 4589–4598 (1995).
20. Liu, J., Wu, T.-C. & Lichten, M. The location and structure of double-strand DNA breaks induced during yeast meiosis: evidence for a covalently linked DNA-protein intermediate. *EMBO J.* **14**, 4599–4608 (1995).
21. Keeney, S. & Kleckner, N. Covalent protein-DNA complexes at 5' strand termini of meiosis-specific double-strand breaks in yeast. *Proc. Natl Acad. Sci. USA* **92**, 11274–11278 (1995).
22. Atcheson, C. L. & Esposito, R. E. Meiotic recombination in yeast. *Curr. Opin. Genet. Dev.* **3**, 736–744 (1993).
23. Cao, L., Alani, E. & Kleckner, N. A pathway for generation and processing of double-strand breaks during meiotic recombination in *S. cerevisiae*. *Cell* **61**, 1089–1101 (1990).
24. Loidl, J., Klein, F. & Shertan, H. Homologous pairing is reduced but not abolished in asynaptic mutants of yeast. *J. Cell Biol.* **125**, 1191–1200 (1994).
25. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–567 (1996).
26. Szent-Gyorgyi, C. A bipartite operator interacts with a heat shock element to mediate early meiotic induction of *Saccharomyces cerevisiae* HSP82. *Mol. Cell. Biol.* **15**, 6754–6769 (1995).
27. Keeney, S., Giroux, C. N. & Kleckner, N. Meiosis-specific DNA double-strand breaks are catalysed by Spo11, a member of a widely conserved protein family. *Cell* **88**, 375–384 (1997).
28. de Massy, B., Baudat, F. & Nicolas, A. Initiation of meiotic recombination in *Saccharomyces cerevisiae* haploid meiosis. *Proc. Natl Acad. Sci. USA* **91**, 11929–11933 (1994).

Acknowledgements. We thank G. R. Smith for communicating the revised *REC12* sequence before publication, R. E. Esposito for *SPO11* and *spo11::URA3* plasmids, B. Labeledan for comments on sequences analysis, F. Lottspeich for peptide sequencing. This work was supported by the CNRS, the Association de Recherche Contre le Cancer, the Ligue Nationale contre le Cancer, the European Community (HCM and Biotech programs), Human Frontier and the MRES (ACC). The sequence of A and B subunit genes has been deposited in EMBL-EBI, accession no. Y10582.

Correspondence and requests for materials should be addressed to P.F. (e-mail: forterre@igmors.u-psud.fr).

Trans-activation of group II intron splicing by nuclear U5 snRNA

Martin Hetzer, Gabriele Wurzer, Rudolf J. Schweyen & Manfred W. Mueller

Vienna Biocenter, Institute of Microbiology and Genetics, University of Vienna, A-1030 Vienna, Austria

Similarities between RNA splicing during autocatalytic excision of group II introns and pre-mRNA processing led to the hypothesis that group II introns might be the evolutionary predecessors of spliceosomal small nuclear RNAs^{1–4}. The ID3 subdomain stem-loop structure of group II introns, the proposed analogue of the spliceosomal U5 snRNA⁵, is thought to be essential for 5' splice site recognition and anchoring of the free 5' exon⁶. Using the group II intron *bII* we have analysed the role of ID3 in splicing. In the absence of ID3 the 5' splice site was recognized accurately and efficiently, but exon anchoring was greatly reduced. This step was restored in the presence of RNA fragments consisting of either the terminal stem-loop structure of ID3 or spliceosomal U5 snRNA. This suggests that the predominant role of both RNAs is to anchor the 5' exon during exon ligation. Furthermore, as U5 complements for the loss of ID3, a similar network of structural RNAs may form the catalytic core of both group II introns and spliceosomes.

Group II introns and nuclear pre-mRNA introns share common features in primary sequence, reaction modes, stereochemical specificities and splicing intermediates^{7,8}. For both types of introns, splicing is initiated by the 2'-hydroxyl group of an adenosine attacking the 5' splice site, followed by exon ligation and release of lariat introns^{9–12}. These similarities supported the idea that fragmentation of group II intron domains (Fig. 1a) might have generated the precursors for the five snRNAs in nuclear splicing². As the potential progenitor of U5 snRNA, group II intron subdomain ID3 was the most likely candidate for a fragmentation assay *in vitro*. The terminal loop of ID3 RNA contains the exon binding site 1 (EBS1) which is complementary to the 3' end of the 5' exon, intron binding site 1 (IBS1) (ref. 13). An additional motif, an α' -site, is located in the internal ID3 loop (Fig. 1a), and seems to be involved in stabilization of the correct folding of the catalytic core¹⁴.

We used an *in vitro* deletion-complementation assay to investigate the role of subdomain ID3 (Fig. 1b) in splicing the group II intron *bII* from yeast mitochondria. First, we analysed the effect of complete ID3 deletion (Fig. 2a). The *bII* wild-type pre-mRNA was processed efficiently into the splicing end products, lariat and the ligated exons. In contrast, mutant *bII*ΔID3 precursors predominantly gave rise to the splicing intermediates, lariat-3' exon and free 5' exon. Only small amounts of end products were observed, indicating a pronounced inhibitory effect of the ID3 deletion on the second splicing step. The efficiency of the first step, branch formation, was only two- to threefold slower than the wild-type reactivity (Fig. 2b). As confirmed by primer extension analysis, the branch-point attack was direct precisely towards the 5' splice site (not shown). This finding was unexpected, as a functional interaction between EBS1 and IBS1 was thought to be required for both efficiency and specificity of 5' splice-site cleavage⁶.

The splicing of complete *bII* pre-RNAs containing mutations in intron binding site 1 had only a minor effect on first-step catalysis (for example, a threefold reduction when wild-type IBS1 motif 5'-GACAGA was changed to 5'-GAATTC; not shown). These findings might seem to contrast with results obtained using group II intron *al5c* (ref. 13), but the *al5c* mutations were tested under low salt conditions, in which splicing is much more sensitive to structural destabilization of the intron than in the high ionic conditions used with the *bII* intron. This could explain at least part of the potential discrepancy.

The severe (80-fold) reduction of the second splicing step of *bII*ΔID3 as compared to the wild-type (Fig. 2c) is consistent with the proposed role of ID3 in preventing premature dissociation of the 5' exon before exon ligation¹³. A similar conclusion has been made regarding the role of U5 snRNA in nuclear pre-mRNA splicing¹⁵.

To investigate whether the severe effect on splicing can be restored in a *trans*-assay, we co-incubated separately transcribed ID3 RNA with the mutant *bII*ΔID3 (Fig. 2a). There was no arrest in the intermediate stage, leading to the excision of lariat RNA and exon ligation. In the presence of excess ID3 RNA the bimolecular splicing system (*bII*ΔID3 + ID3) was as reactive as the monomolecular reactions (wild type) (Fig. 2b, c). Thus association of the two pieces efficiently reconstituted the catalytic core of intron *bII* *in trans*.

A *trans*-assay using a truncated form of ID3 RNA (ID3SL) lacking the α' -site and containing non-canonical stem nucleotides (Fig. 1b) showed that the terminal loop structure of ID3 RNA was sufficient to promote *bII*ΔID3 splicing (Fig. 3). The reduced reactivity rate compared to the assay with the complete ID3 structure may be explained in part by the loss of the interaction between α and α' . The finding that the linear form of the ID3 loop sequences (EBS1) did not stimulate *bII*ΔID3 splicing *in trans* (not shown) indicates that a terminal stem-loop structure containing the EBS1 sequences is crucial for ID3 function.

The relevance of the ID3 loop size for *trans*-activation was tested using two ID3 variants, which had the four loop residues adjacent to

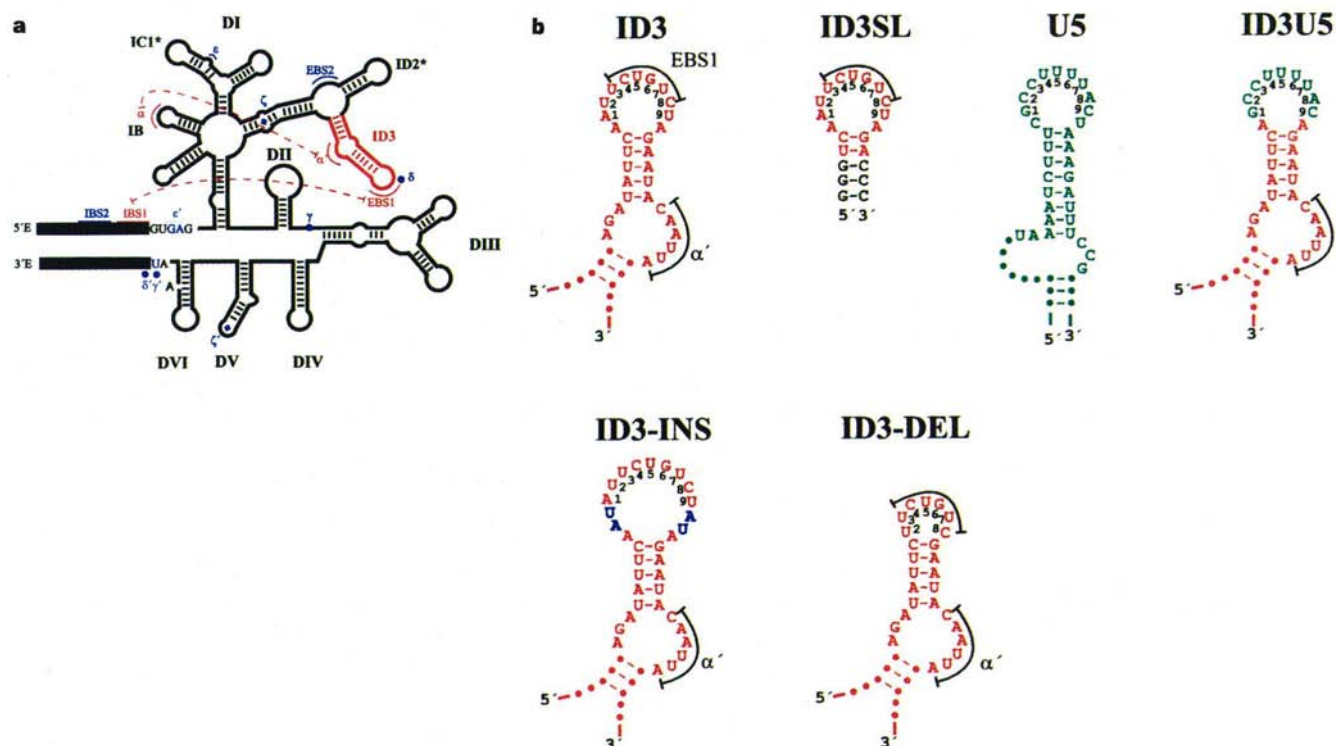


Figure 1 Sequences and secondary structure of group II intron *b11* and trans-acting RNA molecules. **a**, Secondary structure and tertiary interactions¹⁶. Black boxes, exon sequences; continuous line, six major intron domains (DI-DVI); 5'-GUGAG...AU-3', consensus nucleotides at the intron boundaries; red line, ID3 subdomain; broken red lines, interactions of ID3 with intron and exon sequences (EBS1/IBS2 and α - α')^{3,14}; asterisks, subdomain deletions analysed in previous studies^{22,23}; note that the ID2 deletion, reported to have no effect on forward splicing²², was included in this study; blue, individual nucleotides and structures

involved in tertiary interactions (EBS2-IBS2, ϵ - ϵ' , δ - δ' , $(\gamma-\gamma')$ and $(\zeta-\zeta')$ ⁶. For *b11* the δ - δ' interaction (U-U) is not established. **b**, Secondary structure of trans-acting RNAs used for complementation. Subdomain ID3 nucleotides involved in tertiary contacts (EBS1 and α' -site) are marked. ID3SL contains three artificial base-pair substitutions (G-C, black). The identical loop size of ID3 and U5 snRNA (human, green) is indicated by numbers. Nucleotide positions 3-6 (U5) have been shown to form base pairs with 5' and 3' exon sequences^{5,17}. Blue indicates inserted nucleotides in ID3-INS.

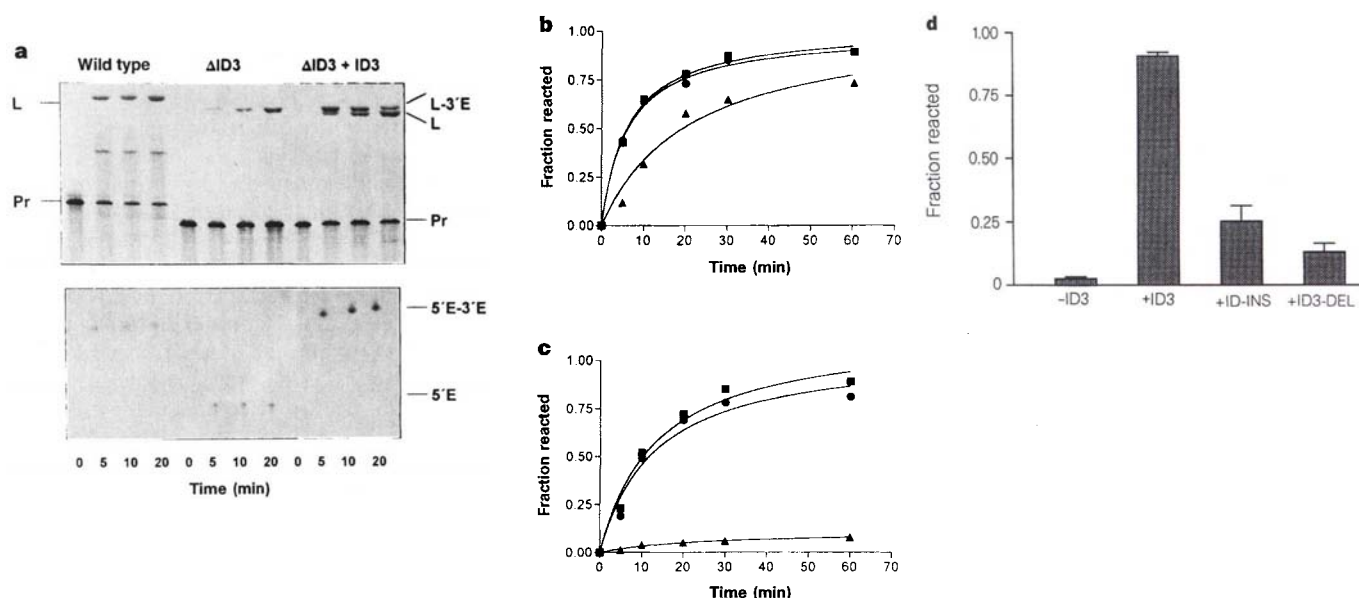


Figure 2 Deletion-complementation assays compared with the self-splicing of wild-type *b11*. **a**, Autoradiogram of the self-splicing reactions of 100 nM *b11* pre-RNAs (ID3 200 nM; multiple turnover conditions). Pr, pre-RNA; L-3'E, lariat-3' exon; 5'E-3'E, ligated exons; 5'E, 5' exon. **b**, Effect of ID3 deletion and complementation on lariat formation. To quantify the first step of splicing, the sum of L-3'E and lariat (L) was divided by this value plus unreacted precursor (P). Wild type (squares), Δ ID3 (triangles), Δ ID3 + ID3 (circles). **c**, Exon ligation (2.0 μ M ID3;

single turnover conditions). The efficiency of the second step was deduced from the ratio of L to L-3'E at the same time point. Additionally, gel-purified L-3'E of *b11* and *b11* Δ ID3 were incubated with saturating 5' exon *in trans*; the efficiency of exon ligation was deduced from the ratio of 5'E-3'E to L-3'E plus 5'E-3'E (not shown). Symbols as **b**, **d**, Effect of ID3 loop size (see Fig. 1b) on trans-activation (60-min values of three independent single-turnover experiments).

the EBS1 motif either deleted (ID3-DEL) or inserted (ID3-INS) (Fig. 1b). Compared with the wild-type ID3 loop (11 nucleotides), the stimulation of second-step catalysis was reduced significantly when ID3-INS or ID3-DEL were added *in trans* (Fig. 2d). Thus a terminal stem-loop structure of defined size containing the EBS1 motif is crucial for *trans*-activation.

The group II intron deletion-complementation assay allowed us to test the proposed spliceosomal counterpart of ID3, namely U5 snRNA, for its ability to substitute functionally for the group II intron structure. As with the *trans*-activation by ID3 RNA, the terminal stem-loop structures of human U5 snRNA restored the autocatalytic excision of *b11ΔID3* lariat RNA (Fig. 4a). Second-step catalysis was increased tenfold, but there was no significant increase in branch formation (Fig. 4b, c); minor amounts of aberrant ligation products were also observed (not shown). A very similar effect of *trans*-activation was obtained when a chimaeric RNA containing the U5 loop sequences and the complete ID3 stem region (ID3U5) (Fig. 1b) was used (not shown).

We then analysed the proposed interaction of the terminal loop of U5 snRNA with the 5' exon of group II intron *b11*. The IBS1 motif within the 5' exon was made strictly complementary to the U5 snRNA loop (*b11ΔID3U5-IBS1*; 5'-UAAAAG). Compared with wild-type *b11ΔID3* splicing, efficient splicing rescue of this mutant was observed when the U5 snRNA loop was added *in trans* (Fig. 4d). This indicated that U5 loop sequences contact the IBS1 through a direct base-pairing interaction. This was supported by experiments using compensatory base-pair exchanges. The four central U residues in the conserved U5 loop were substituted by adenosines (ID3U5K), and the four adenosines in *b11ΔID3U5-IBS1*

(5'UAAAAG) were changed to uridines (U5K-IBS1; UUUUUG). The combinations that destroy a direct base-pairing interaction between the U5 loop and the IBS1 site failed to stimulate splicing (Fig. 4d). The compensatory combinations, however, efficiently restored the exon ligation step. From this we conclude that restoration of second-step catalysis is governed by a terminal loop structure and by the potential of the loop sequences to anchor the exon sequences, in which the central U residues of U5 snRNA are involved in this interaction.

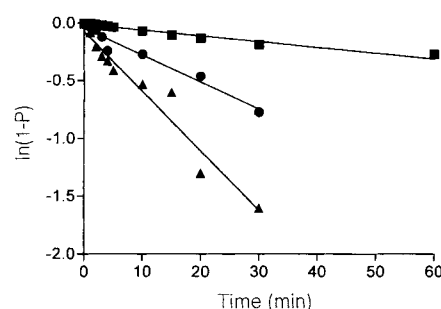
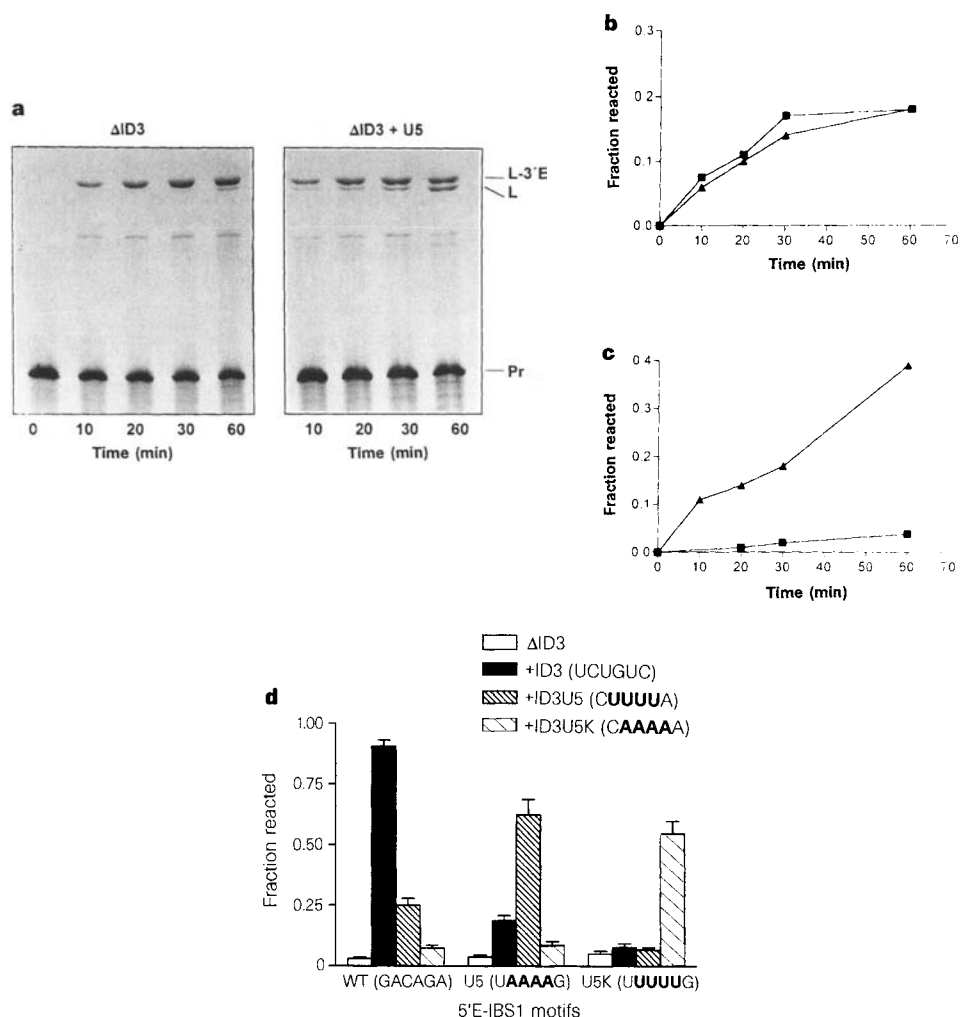


Figure 3 Comparison of reaction time courses of $\Delta ID3$ (squares) with the *trans*-acting RNAs ID3 (triangles) and ID3SL (circles) under single-turnover conditions. In the self-splicing reactions, 50 nM $\Delta ID3$ pre-RNA were incubated with 2.5 μ M ID3 and ID3SL RNA, respectively. Data were plotted as the natural logarithm of the fraction of precursor RNA remaining during the reaction time (P is the fraction of reaction products).

Figure 4 *Trans*-activation of $\Delta ID3$ self-splicing by U5 snRNA structures. **a**, Self-splicing of *b11ΔID3* precursors (100 nM) in the absence and presence of human U5 snRNA (5.0 μ M). Only the conserved terminal stem-loop structure (see Fig. 1b) was assayed, as most of U5 sequences are not required for its function²⁴. **b, c**, Effect of U5 snRNA addition (triangles) to $\Delta ID3$ alone (squares) on first-step (**b**) and second-step (**c**) catalysis. **d**, Importance of base-pairing interactions for second-step restoration using different IBS1 and loop sequences. The IBS1 mutants are indicated on the x-axis. Efficiency of the *trans*-activation was calculated as described (Fig. 2b-d).



In nearly all group II introns, contacts between ID3 and exons are established as specific Watson–Crick base-pairing interactions (IBS–EBS)¹⁶, whereas the U5 snRNA loops bind to different exon sequences. It has been shown that U5 snRNA with drastically altered loop sequences could still carry out both splicing steps¹⁵. As can be deduced from Fig. 4d, alteration of loop sequences in group II intron splicing severely reduced *trans*-activation, and therefore demonstrates that there is a substantial requirement for base-pairing interactions. Our data suggest that the ability of the uridine-rich U5 snRNA loop to bind exons promiscuously¹⁷ may enable this structure to *trans*-activate not only *bII*, but also some other autocatalytic group II splicing systems.

In conclusion, our findings suggest that group II intron sub-domain ID3 and spliceosomal U5 snRNA are functional counterparts. In both splicing systems, the corresponding structures are dispensable for the accuracy of 5' splice-site cleavage, but are required to align the 5' and 3' exons for ligation. This suggests that a similar catalytic structure is formed by a network of *trans*- or *cis*-acting RNAs, generating a pocket for anchoring either ID3 or U5 snRNA to contact the exon–intron boundaries. However, there is an important difference in the nature of the loop–exon interaction. ID3 RNA requires a Watson–Crick base-pairing interaction, whereas U5 snRNA contacts the exons in a non-Watson–Crick manner¹⁸ coordinated by as yet unidentified factors. □

Methods

BS/*bII*ΔID3 was generated from BS/*bII* (ref. 19) using two partly complementary oligonucleotides (5'-AAAATATCGTAACATAACAATATATTCGGTATAAAATTAGAAATCC-3' and 5'-GGATTCTAATTTTATACCGAATATATTGTTATGTTACGATATTTT-3') in two separate polymerase chain reactions. A second PCR (5'-AAATCTGGTACCATGGCTGGGAACAAAAGGTTATGTGTGTG-3'; 5'-GTCTGAATCCATGGCTGCAATTGTGATAGGTAGATC-3') amplified *bII*ΔID3, which was subsequently ligated into BSKS+ through *KpnI* and *EcoRI*. Pre-mRNAs were synthesized by transcription of *EcoRI*-digested BS/*bII* and BS/*bII*ΔID3 with T3 RNA polymerase in the presence of ³⁵S-UTP followed by gel purification²⁰. Small *trans*-acting RNA molecules were generated by PCR using primers with 12-nucleotide complementary regions at their 3' end (plus primer containing the sequence for T7 promoter) followed by *in vitro* transcription and gel purification. Before incubation, RNA molecules were heated to 95 °C for 2 min and then cooled to 37 °C (ref. 21). The pretreated samples were incubated for 0, 5, 10 or 20 min in a buffer containing 40 mM Tris-HCl, pH 7.5, 60 mM MgCl₂, 5 mM spermidine, and 1.25 M NH₄Cl (Fig. 2) or 0.5 M NH₄Cl (Figs 3, 4) at 37 °C. Self-splicing reaction products were

separated on 5% polyacrylamide–8 M urea gels. The rate of turnover was quantified with the Molecular Dynamics phosphorimager.

Received 3 December 1996; accepted 31 January 1997.

1. Jacquier, A. Self-splicing group II introns and nuclear pre-messenger RNA introns—how similar are they? *Trends Biochem. Sci.* **15**, 351–354 (1990).
2. Sharp, P. A. Five easy pieces. *Science* **262**, 663 (1991).
3. Weiner, A. M. mRNA splicing and autocatalytic introns: distant cousins or the products of chemical determinism? *Cell* **72**, 161–164 (1993).
4. Wise, J. A. Guides to the heart of the spliceosome. *Science* **262**, 1978–1979 (1993).
5. Newman, A. J. & Norman, C. U5 snRNA interacts with exon sequences at 5' and 3' splice-sites. *Cell* **68**, 743–754 (1992).
6. Michel, F. & Ferat, J.-L. Structure and activities of group II introns. *Annu. Rev. Biochem.* **64**, 435–461 (1994).
7. Padgett, R. A., Podar, M., Boulanger, S. C. & Perlman, P. S. The stereochemical course of group II intron self-splicing. *Science* **266**, 1685–1688 (1994).
8. Moore, M. J. & Sharp, P. A. Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. *Nature* **365**, 364–368 (1993).
9. Peebles, C. L., Perlman, P. S., Mecklenburg, K. L., Petrillo, M. L. & Tabor, J. H. A self-splicing RNA excises an intron lariat. *Cell* **44**, 213–223 (1986).
10. Schmelzer, C. & Schwyen, R. J. A self-splicing of group II introns *in vitro*: mapping of the branch point and mutational inhibition of lariat formation. *Cell* **46**, 557–565 (1986).
11. Van der Veen, R. *et al.* Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing *in vitro*. *Cell* **44**, 225–234 (1986).
12. Moore, M. J., Query, C. C. & Sharp, P. A. in *Tech RNA World* (eds Gesteland, R. F. & Atkins, J. E.) 303–358 (Cold Spring Harbor Laboratory Press, NY, 1993).
13. Jacquier, A. & Michel, F. Multiple exon-binding sites in class II self-splicing introns. *Cell* **50**, 17–29 (1987).
14. Harris-Kerr, C. L., Zhang, M. & Peebles, C. L. The phylogenetically predicted base-pairing interaction between α and α' is required for group II splicing *in vitro*. *Proc. Natl. Acad. Sci. USA* **90**, 10658–10662 (1993).
15. O'Keefe, R. T., Norman, C. & Newman, A. J. The invariant U5 snRNA loop I sequence is dispensable for the first catalytic step of pre-mRNA splicing in yeast. *Cell* **86**, 679–689 (1996).
16. Michel, F., Umesono, K. & Ozeki, H. Comparative and functional anatomy of group II catalytic introns—a review. *Gene* **82**, 5–30 (1989).
17. Madhani, H. D. & Guthrie, C. Dynamic RNA–RNA interactions in the spliceosome. *Annu. Rev. Genet.* **28**, 1–26 (1994).
18. Sontheimer, E. J. & Steitz, J. A. The U5 and U6 small nuclear RNAs as active components of the spliceosome. *Science* **262**, 1989–1996 (1993).
19. Mueller, M. W., Stocker, P., Hetzer, M. & Schwyen, R. J. The fate of the junction phosphate in alternating forward and reverse splicing reactions of group II intron RNA. *J. Mol. Biol.* **222**, 145–154 (1991).
20. Mueller, M. W., Hetzer, M. & Schwyen, R. J. Group II intron RNA catalyzes progressive, site-specific, 3' to 5' nucleotide insertion: a model for RNA editing? *Science* **261**, 1035–1038 (1993).
21. Michels, J. W. & Pyle, A. M. Conversion of a group II intron into a multiple-turnover ribozyme that selectively cleaves oligonucleotides: elucidation of reaction mechanism and structure/function relationships. *Biochemistry* **34**, 2965–2977 (1995).
22. Bachl, J. & Schmelzer, C. Effects of deletions at structural domains of group II intron *bII* on self-splicing introns. *J. Mol. Biol.* **212**, 113–125 (1990).
23. Suchy, M. & Schmelzer, C. Restoration of the self-splicing activity of a defective group II intron by a small *trans*-acting RNA. *J. Mol. Biol.* **222**, 179–187 (1991).
24. Frank, D. N., Roiha, H. & Guthrie, C. Architecture of the U5 small nuclear RNA. *Mol. Cell. Biol.* **14**, 2180–2190 (1994).

Acknowledgements. This work was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (F.W.F.).

Correspondence and requests for materials should be addressed to M.W.M. (e-mail: mueller@gem.univie.ac.at) or M.H. (e-mail: hetzer@gem.univie.ac.at).

YOURS TO HAVE AND TO HOLD BUT NOT TO COPY

The publication you are reading is protected by copyright law.

Photocopying copyright material without permission is no different from stealing a magazine from a newsagent, only it doesn't seem like theft.

If you take photocopies from books, magazines and periodicals at work your employer should be licensed with CLA.

Make sure you are protected by a photocopying licence.



The Copyright Licensing Agency Limited
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931 Fax: 0171 436 3986